UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

Lucas Guedes de Pádua

Análise da composição da microbiota intestinal associada a um estado de doença utilizando aprendizado de máquina

Ribeirão Preto
2022

Lucas Guedes de Pádua

Análise da composição da microbiota intestinal associada a um estado de doença utilizando aprendizado de máquina

Trabalho de Conclusão de Curso apresentado ao Curso de Informática Biomédica, da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Bacharel em Informática Biomédica.

Orientador: Profa. Dra. María Eugenia Guazzaroni

Este trabalho foi apresentado e aprovado pela Comissão Coordenadora do Curso deInformática Biomédica em 15/03/2023.

Ribeirão Preto

2022

FICHA CATALOGRÁFICA

Pádua, Lucas Guedes de

Análise da composição da microbiota intestinal associada a um estado de doença utilizando aprendizado de máquina.

Trabalho de Conclusão de Curso apresentado ao Curso de Informática Biomédica, da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Bacharel em Informática Biomédica.

Orientadora: Guazzaroni, Maria Eugenia.

1. Microbiota intestinal; 2. Aprendizado de máquina; 3. Doenças sistêmicas.

Dedico esse trabalho aos meus pais, José e Alessandra, que sempre me apoiaram e acreditaram em mim.

AGRADECIMENTOS

À Universidade de São Paulo e à Faculdade de Medicina de Ribeirão Preto pelo ensino e pelas oportunidades.

À todos os professores que tive durante o meu tempo de graduação pelos ensinamentos e experiências.

À Profa. Dra. Maria Eugênia Guazzaroni pela oportunidade, atenção e orientação no desenvolvimento deste trabalho.

Ao Dr. Rafael Silva Rocha, A Dra. Stela Virgilio pela oportunidade que me deram e pela confiança depositada em mim.

Ao Gustavo Tamasco pela companhia, ajuda e ensinamentos durante o meu estágio.

Aos meus colegas da Turma 17 do Curso de Informática Biomédica. Vou levar sempre com muito carinho as memórias de tudo que se passou durante esses quatro anos, foi um prazer dividir essa experiência com vocês.

Aos meus amigos Victor Malheiro, Julia Saito, Giullia Inoue, Marina Priolo, Alexsandra Beatriz, João Pedro, Ítalo Botura e Felipe Marcelo pela vivência e fraternidade. Vocês fizeram esses quatro anos serem inesquecíveis e muito melhores.

Aos meus pais José Domingos e Alessandra Guedes, pelo apoio, ajuda, sacrifícios, amor e carinho por mim. Definitivamente se não fosse por vocês eu não estaria aqui hoje.

RESUMO

A microbiota intestinal é o conjunto de microrganismos que habitam o trato intestinal. Tais microrganismos exercem funções essenciais no metabolismo, no sistema imune e na nutrição. A composição da microbiota traz informações importantes sobre o estado de saúde de um indivíduo. A presença de uma determinada doença, altera o estado saudável do corpo, e essa alteração pode ser vista através da distribuição das espécies presentes no intestino. Por esse motivo, olhar um certo fenótipo através da composição da microbiota pode vir a desvendar mecanismos e características novas da doença, e auxiliar na prevenção, diagnóstico e tratamento dessas condições. Levando em conta a crescente na geração de dados de larga escala e evolução do processamento computacional, o uso do aprendizado de máquina na análise da microbiota é uma ferramenta eficaz na hora de associar a composição com a aparição de uma doença. Neste trabalho foram coletados dados da microbiota intestinal de pessoas saudáveis e de pessoas com obesidade e Diabetes Tipo 2. Seguidamente, dois algoritmos de aprendizado de máquina foram utilizados, Regressão Logística e *Random Forest*, para predição da doença e busca de biomarcadores. A Regressão Logística obteve melhores resultados de classificação do que o algoritmo de *Random Forest*, e ambos encontraram diferentes marcadores para as duas doenças.

Palavras-chave: Microbiota intestinal; Aprendizado de máquina; *Random Forest*; Regressão Logística; Obesidade; Diabetes.

ABSTRACT

The gut microbiota is the collection of all microorganisms that inhabit the Gastro-Intestinal Tract. These microorganisms play essential roles in our metabolism, immune system and nutrition. The microbiota composition brings important information about the health state of an individual. The presence of a disease alters the health state of the body, and this alteration can be seen through the distribution of the microorganism's species in our intestine. For this reason, having a look at a phenotype through the microbiota composition can unveil new mechanisms and characteristics of the disease. Taking the increase in data generation and the evolution of computational processing into account, the use of machine learning on the microbiota analysis is an effective tool when associating the composition with disease onset. In this work, data of the gut microbiota from healthy people and from people with obesity and Type 2 Diabetes was collected. Then, two machine learning algorithms were used, Logistic Regression and Random Forest, for disease prediction and biomarker search. The Logistic Regression got better classification results than the Random Forest, and both found different biomarkers for both diseases.

Key words: Gut microbiota; Machine learning; *Random Forest*; Logistic Regression; Obesity; Diabetes.

SUMÁRIO

1. INTRODUÇÃO

1.1. Problemática abordada

A microbiota intestinal se trata do conjunto de microrganismos que habitam o trato gastrointestinal. Ela é composta por mais de 1000 espécies e está envolvida em diversas funções essenciais como no metabolismo e na nutrição (SEKIROV et al., 2010). A composição da microbiota, ou seja, quais espécies estão presentes no ambiente do trato, tem uma relação direta com o estado fisiológico do hospedeiro, e mudanças nessa composição estão relacionadas às doenças de obesidade (LEY et al., 2006) e Diabetes Tipo 2 (QIN et al., 2012). Dada essa associação, buscar determinadas espécies que caracterizam essas doenças e usar a composição da microbiota intestinal como preditivo da manifestação delas é de grande valor.

1.2. Avanço nas tecnologias de sequenciamento

Com o decorrer dos anos, a partir da descoberta da estrutura do DNA, as tecnologias de sequenciamento, e de análise computacional avançaram, e tal avanço permitiu a descoberta de todo um novo universo no corpo humano, invisível ao olho nu, os microrganismos. Desde o seu início as tecnologias de sequenciamento progrediram muito, começando com o método de Sanger criado em 1977 (ALVES et al., 2018), que possibilitou posteriormente o primeiro sequenciamento inteiro do genoma humano. Sequenciar era um processo que exigia bastante tempo e dinheiro, contudo, com o progresso do sequenciamento, o custo e tempo demandados diminuíram significativamente (ALVES et al., 2018). O Projeto Genoma Humano levou 15 anos e aproximadamente 100 milhões de dólares para ser concluído usando o sequenciamento Sanger, enquanto o mesmo sequenciamento usando o 454, uma tecnologia de sequenciamento da segunda geração, demorou apenas 2 meses e custou um centésimo do valor (WHEELER et al., 2008).

A possibilidade de sequenciar genomas em menos tempo e gastando menos dinheiro permitiu o surgimento de novas áreas do conhecimento, que usufruem das vantagens principalmente de sequenciadores, agora de terceira geração, que descartam a necessidade de amplificação das amostras por PCR e conseguem sequenciar leituras mais longas, mais rapidamente, facilitando posteriormente a montagem do genoma final, e tudo isso com um custo ainda mais baixo que os sequenciadores da geração anterior (KCHOUK et al., 2017).

Os sequenciadores de nova geração possibilitaram a criação e crescimento da Metagenômica, que visa estudar os genomas de todos os microrganismos presentes em uma comunidade. Enquanto uma comunidade de microrganismos é chamada de microbiota, o conjunto de genomas desses microrganismos é denominado de microbioma. A vantagem que os sequenciadores de nova geração oferecem permite que todas as espécies presentes em uma microbiota tenham seu genoma representado de forma significativa, mesmo quando em abundâncias diferentes (ROUMPEKA et al., 2017). A Metagenômica é de extrema importância, pois através dela é possível estudar microrganismos dificilmente cultiváveis, analisando seus genomas em uma determinada comunidade (ALVES et al., 2018).

Um exemplo de sequenciador de terceira geração é o MinION da Oxford Nanopore. O MinION é um dispositivo portátil que mede aproximadamente 10 centímetros, e se comunica com o computador por meio de uma entrada USB 3.0. O sequenciamento acontece através de nanoporos, que são orifícios na nanoescala construídos com proteínas ou material sintético. As duas fitas do DNA passam pelo nanoporo em sequência, e a mudança dos nucleotídeos enquanto as fitas passam pelo nanoporo causam uma mudança na corrente iônica, que é captada e colocada em um gráfico, onde cada corrente irá representar uma base (KCHOUK et al., 2017). A Figura 1 mostra um exemplo do minION conectado a um computador.

Figura 1. Sequenciador minION conectado em um notebook. Retirado de: https://www.genengnews.com/insights/first-nanopore-sequencing-of-human-genome/

1.3. Função e importância da microbiota intestinal

Através do sequenciamento de nova geração e das ferramentas mais avançadas de bioinformática, é possível explorar a diversidade genética de uma comunidade não-cultivável (SHARPTON., 2014). Usando esse método foi realizada a caracterização da microbiota humana, o conjunto de microrganismos presentes no corpo humano.

A microbiota humana é o conjunto de microrganismos que habita diferentes partes do corpo. Nele se encontram 10 vezes mais microrganismos do que células humanas, e 100 vezes mais quando comparado o número de genes humanos codificados com o número de genes microbianos codificados. A microbiota intestinal é composta por mais de 1000 espécies a nível de filotipo (LOZUPONE et al., 2012; CLAESSON et al., 2009), e está envolvida no metabolismo, na função imunológica, fisiologia e nutrição.

Estando presente em todas as partes do corpo, no trato intestinal é onde é encontrada a maior diversidade microbiana no ser humano, mesmo sendo composta por mais de 1000 espécies, a microbiota intestinal se resume majoritariamente em apenas 2 filos: os Firmicutes e os Bacteroidetes (SEKIROV et al., 2010). Essa diversidade de espécies se dá por conta da redundância funcional na microbiota. Pessoas diferentes divergem muito na composição dos microrganismos que habitam o seu trato

gastrointestinal, contudo, os perfis de genes funcionais são bem mais similares (LOZUPONE et al., 2012), isso porque diversas funções foram transferidas entre espécies, não necessariamente relacionadas filogeneticamente, no decorrer do tempo. O anterior garante que caso haja uma diminuição da presença de uma certa espécie, possivelmente uma outra espécie possa assumir as tarefas que a anterior exercia, garantindo assim a manutenção de um estado saudável (LOZUPONE et al., 2012).

A composição da microbiota tem uma relação direta com o estado fisiológico do hospedeiro, e algumas mudanças nessa composição contribuem diretamente para um estado de doença. Diversos fatores podem influenciar na alteração e possível disrupção da microbiota intestinal, como mudanças na dieta, estresse, idade, ambiente, assim como fatores genéticos. A alteração da composição da microbiota pode ser associada a doenças como obesidade (LEY et al., 2006), COVID-19 (ZUO et al., 2020), câncer de pâncreas (Kartal et al., 2022) e Diabetes Tipo 2 (QIN et al., 2012). Em certas doenças, as divergências são consistentes entre diferentes indivíduos, quando comparados a indivíduos saudáveis, como na doença de Chron (WILING et al., 2010), porém em outros casos existe divergência dos grupos controle, mas não há consistência entre os grupos estudados (CHANG et al., 2008).

1.3.1. A microbiota intestinal da obesidade

A obesidade, descrita como a epidemia do século 21 segundo a Organização Mundial da Saúde (OMS), é uma doença multifatorial, o que significa que existem várias causas que produzem sobrepeso e obesidade, e não apenas uma. Dentre elas podemos destacar as causas genéticas, metabólicas, psicológicas, socioculturais, neuroendócrinas, sedentarismo, microbiota intestinal e alimentação hipercalórica. A prevalência da obesidade no mundo quase triplicou desde 1975, sendo que no início se tratava apenas de um problema em países desenvolvidos. Com o decorrer dos anos essa condição começou a se espalhar por países em desenvolvimento, e em 2016, 13% da população mundial adulta era obesa (OMS, 2021). As principais causas para esse aumento alarmante do nível de obesidade, podem ser associadas à crescente urbanização no último século, e maior consumo de alimentos gordurosos. A obesidade é a causa de diversas mortes ao redor do mundo, além de aumentar o risco de doenças como diabetes, doenças cardiovasculares, condições músculo-esqueletais e câncer (OMS, 2021). Sendo uma das consequências da

má nutrição, a obesidade é uma doença que pode ser prevenida com reeducação de hábitos alimentares e prática de exercícios físicos (OMS, 2021).

Mesmo que a obesidade sendo considerada uma doença multifatorial, diversos estudos buscaram entender a fundo o envolvimento da microbiota intestinal e sua composição na obesidade. Fatores como a diminuição da diversidade da microbiota, razão entre a abundância de Firmicutes e Bacteroidetes, além da diferença em abundância de gêneros de bactérias responsáveis por funções metabólicas foram apontados como possíveis alterações decorrentes da obesidade. Porém a inconsistência desses achados em diferentes estudos e dificuldade de entender a relação de causalidade entre essas características e a obesidade, impõem um desafio no entendimento da relação entre a microbiota e a obesidade (Tagliabue, 2013).

1.3.2. A microbiota intestinal da diabetes

A diabetes é uma doença sistêmica, causada pela inabilidade do corpo humano de produzir insulina suficiente, ou de conseguir utilizá-la de forma eficaz. O número de pessoas diabéticas no mundo saltou de 108 milhões em 1980 para 480 milhões em 2014. A diabetes tipo 2, tipo mais comum, compondo 95% de todos os casos de diabetes, é consequência direta do sobrepeso e da inatividade física. A diabetes aumenta o risco de condições no rim, de doenças cardiovasculares, de amputação de membros e cegueira, e foi causa de 2 milhões de mortes apenas em 2019. A diabetes também é uma doença que pode ser evitada com bons hábitos alimentares e prática de exercícios físicos (OMS, 2022).

Em estudos sobre a microbiota intestinal de pessoas diabéticas, uma disbiose pôde ser observada, com diminuição da abundância de bactérias benéficas, principalmente aquelas envolvidas em produzir ácido butírico e aumento em bactérias conhecidas por serem oportunistas. No entanto, assim como nos estudos sobre obesidade, a diversidade da abundância dessas bactérias oportunistas foi alta, e dificulta o entendimento da participação delas no quadro de Diabetes (QIN et al., 2012).

Como demonstrado anteriormente, obesidade e diabetes tipo 2 são doenças altamente relacionadas, com causas, riscos e formas de prevenção semelhantes, causando

milhares de mortes anualmente, e a sua presença só aumenta cada vez mais, consequência de um mundo cada vez mais urbanizado, com comidas cada vez mais industrializadas e aumento de hábitos não saudáveis como pouca atividade física (OMS, 2022). A obesidade pode ainda ser uma das causas de desenvolvimento da diabetes (CHATTERJEE; GERDES; MARTINEZ, 2020). Desvendar, por meio da microbiota intestinal, as características que podem levar a prevenção, diagnóstico ou terapia dessas doenças é essencial para entender o funcionamento complexo dessas condições, e impedir que ainda mais pessoas sejam acometidas por elas, num futuro onde caminhamos para ainda mais sedentarismo e uma alimentação cada vez menos balanceada**.**

1.4. Uso do aprendizado de máquina no reconhecimento de padrões

O aprendizado de máquina é a capacidade de um computador aprender sem ser explicitamente programado (SAMUEL., 2000). Usando desse aprendizado, o computador é capaz de gerar uma saída, ou uma previsão, baseado em dados de entrada inéditos. O aprendizado de máquina pode ser separado em diversas áreas, destacando duas delas, o aprendizado não supervisionado e o aprendizado supervisionado (Figura 2).

O aprendizado não supervisionado permite analisar e categorizar conjuntos de dados não classificados, onde não há uma saída previamente conhecida. Tal técnica viabiliza a exploração e extração de informações quando pouco se sabe sobre como os dados interagem entre si. Esse tipo de aprendizado é mais comumente usado quando o objetivo é de agrupar entradas semelhantes ou reduzir a dimensionalidade dos dados. O aprendizado supervisionado usa de conjuntos de dados que possuem tanto as entradas, como as saídas, e usando dessas informações aproxima uma função que melhor descreve esses dados, a usando posteriormente para fazer predição em entradas inéditas, que não se conhece a saída, ou a resposta. Um dos exemplos de tarefas do aprendizado supervisionado é a classificação. (MAHESH, 2018).

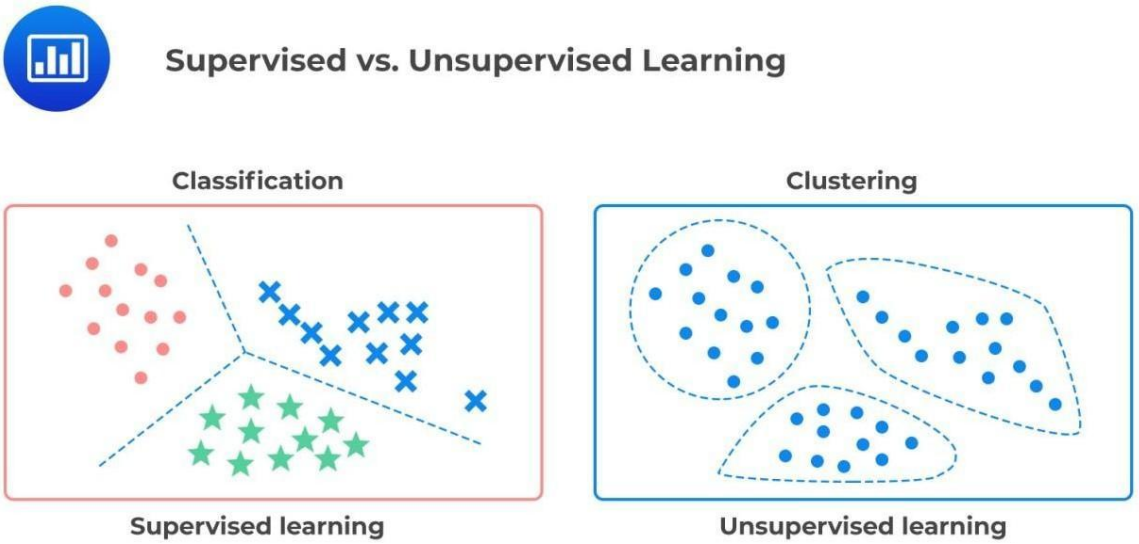


Figura 2: Na esquerda, uma tarefa típica do aprendizado supervisionado, a classificação, e na direita uma típica do aprendizado não supervisionado, o agrupamento, ou clusterização. Retirado de: https://www.linkedin.com/pulse/supervised-vs-unsupervised-learning-whats-difference-smriti-sai ni/

A grande vantagem do uso de algoritmos de aprendizado de máquina se demonstra quando estamos trabalhando com dados de alta dimensionalidade e complexidade. Quando nós temos dados com essas características, se torna praticamente impossível encontrar e extrair informações deles usando métodos convencionais (MAHESH, 2018). Isso ocorre quando existem muitas variáveis, na sua maioria na casa das centenas, e todas elas contribuem para a resposta final, dificultando em muito a interpretabilidade dos dados. Com o aumento significativo da coleta e disponibilização de dados, o uso de técnicas de aprendizado de máquina se torna cada vez mais presente e fundamental na extração de informações de dados complexos, sendo essencial para o progresso tecnológico (AKINSOLA, 2017).

1.4.1. O funcionamento do aprendizado de máquina supervisionado

O aprendizado de máquina supervisionado é um método que faz o caminho inverso. De fato, nunca será possível saber qual a função que perfeitamente descreve os nossos dados. Por meio de um método iterativo, ou seja, uma operação repetida diversas vezes, e utilizando um conjunto de dados onde se têm os valores de entrada e saída, é possível aproximar os parâmetros de uma função hipotética que melhor descreve aqueles dados, parafraseando. Ou seja, o algoritmo permite que se aprenda dos dados, para que em uma

entrada de dados inédita, onde não se sabe a saída, utilizando da função aproximada e dos parâmetros aprendidos, seja possível fazer uma predição do resultado final (DHAR et al., 2020).

Existem diversos algoritmos de aprendizado de máquina supervisionado, e cada um deles possui características próprias, como funções de minimização de erro, tipo de predição, estrutura, complexidade, interpretabilidade, entre outras. Cada algoritmo tem sua peculiaridade, e sua aplicação terá um desempenho diferente, dependendo das características do conjunto de dados de interesse (SINGH; THAKUR; SHARMA, 2016). Em suma, o que é oferecido por cada algoritmo diferente, é um conjunto de hipóteses, e quando o modelo aprende dos dados, ele encontra nesse conjunto de hipóteses, uma que melhor os descreve. Quando falamos em uma hipótese que melhor descreve os dados, queremos dizer que quando essa hipótese foi usada pelo modelo para predizer as saídas, ela gerou o menor erro possível (ABU-MOSTAFA, 2012).

Assim, o aprendizado supervisionado é o método de preferência quando se tem conhecimento sobre a estrutura dos dados, e se busca encontrar padrões que possam explicar a distribuição deles, permitindo que o algoritmo infira uma saída utilizando das informações descobertas, quando frente a uma entrada desconhecida pela máquina (MAHESH, 2018).

1.4.1.1. Regressão Logística

A regressão logística é um método de modelagem matemática, que permite associar diversas variáveis a uma saída dicotômica, ou seja, uma saída que pode apenas assumir dois valores. A regressão logística é um modelo muito utilizado pela sua robustez, simplicidade e também pela característica da saída que ele retorna. O modelo logístico usa de uma função que sempre retorna um valor entre 0 e 1, fazendo com que fique fácil a interpretabilidade do resultado. Esse tipo de saída também permite que sejam ajustados os limites para quando assumir se uma saída vai ser 0 ou 1, alterando o valor mínimo necessário da probabilidade para que as classes sejam escolhidas (KLEINBAUM; KLEIN, 2010).

1.4.1.1.1. Regularização LASSO

Uma das limitações da regressão logística é o risco de um modelo enviesado, principalmente quando o número de variáveis é superior ou muito próximo ao número de observações, fazendo com que o modelo não tenha muitos exemplos para aprender. A regularização LASSO (*Least Absolute Shrinkage and Selection Operator)* é uma das abordagens para esse problema. Essa regularização impõe um limite, um valor máximo, que a soma de todos os valores absolutos dos coeficientes aproximados pode assumir. Isso permite que o algoritmo diminua, e até mesmo zere, o valor de alguns coeficientes que não contribuem para a minimização do erro (RANSTAM; COOK, 2018).

Essa limitação diminui a complexidade do modelo e melhora a sua performance, visto que o modelo construído fica mais generalizado. Isso ocorre, pois as chances de o modelo estar enviesado diminuem consideravelmente por causa do menor número de variáveis, resultando em um modelo cuja distribuição não se ajustou ao ruído presente nos dados. A regularização também permite a visualização das variáveis mais importantes, que mais contribuíram para o modelo final (RANSTAM; COOK, 2018).

1.4.2. Random Forest

Random Forest, ou floresta aleatória, é um método de *ensemble*, que usa árvores de decisão como o modelo base. Métodos de *ensemble* usam vários modelos simultaneamente para construir um modelo final e fazer a predição (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). No caso da *Random Forest*, várias árvores de decisão são treinadas, usando diferentes frações dos dados e das variáveis. Para fazer uma nova predição, um dado inédito vai ser analisado por todas as árvores geradas, e no caso da classificação, a categoria que mais apareceu entre as árvores vai ser escolhida. O método de *Random Forest* impede que os resultados sejam enviesados, uma característica conhecida das classificações usando árvores de decisão, devido ao uso de apenas uma fração dos dados conhecidos, e seleção apenas das árvores que deram os melhores resultados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Por causa dessa seleção de árvores, as *Random Forest* conseguem calcular quais variáveis geraram os melhores resultados, e é possível visualizar essas variáveis posteriormente, extraindo mais informações ainda do conjunto de dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

1.4.3. O uso de métodos de aprendizado de máquina em estudos de microbiota intestinal

Na última década, houve um aumento significativo no número de estudos relacionados ao estado da microbiota intestinal associado com diversas doenças, e também de catalogações massivas da microbiota saudável, como o *Human Microbiome Project* (HUTTENHOWER et al., 2012), que por sua vez aumentou a quantidade de dados gerados, e disponíveis sobre a microbiota intestinal humana (MARCOS-ZAMBRANO et al., 2021).

O aprendizado de máquina foi e é usado extensivamente na bioinformática, e para vários fins, como na avaliação de diferentes métodos de classificação (STATNIKOV et al., 2005), ou generalização de modelos em diferentes estudos (PASOLLI et al., 2016). Também, o uso de algoritmos de aprendizado de máquina pode trazer novas descobertas em relação a composição e estrutura da microbiota e características de determinados fenótipos, além da descoberta de biomarcadores dessas doenças. Tais algoritmos podem ainda ser usados na medicina personalizada, para tratamentos mais eficazes e ainda na predição da aparição de uma condição, usando os dados da microbiota intestinal de um indivíduo como parte do diagnóstico (MARCOS-ZAMBRANO et al., 2021).

1.5. A linguagem de programação Julia

Julia é uma linguagem de programação compilada, e de tipagem dinâmica, desenvolvida desde o início com enfoque na performance, porém com a facilidade de uma sintaxe simples e compreensível. A linguagem de programação Julia visa resolver o problema da necessidade de usar duas linguagens quando a performance e eficiência de um algoritmo é necessária, porém o algoritmo foi escrito primariamente em uma linguagem dinâmica de alto nível (BEZANSON et al., 2012).

Isso ocorre porque é muito mais fácil escrever algoritmos em linguagens dinâmicas de alto nível, pela simplicidade da sintaxe e proximidade com a linguagem humana. Porém essas linguagens de programação deixam a desejar na performance, acarretando na reescrita do algoritmo em uma linguagem de baixo nível quando é necessário otimizar o desempenho do programa (BEZANSON et al., 2012). Julia então consegue resolver esse problema, tendo a performance de uma linguagem compilada

estaticamente, porém com a flexibilidade e produtividade de linguagem interpretada de alto nível. A linguagem consegue esse feito devido a tecnologias únicas como despacho múltiplo, e compilação *Just In Time*, que agilizam a execução do programa (BEZANSON, et al., 2012).

2. OBJETIVOS

O presente trabalho tem como objetivo analisar dados de abundância relativa de microrganismos da microbiota intestinal, de pessoas saudáveis e pessoas que possuem obesidade ou diabetes tipo 2, utilizando aprendizado de máquina, a fim de construir um modelo preditivo desses fenótipos e buscar biomarcadores que caracterizem a microbiota intestinal associada a essas doenças.

2.1. Objetivos específicos

1) Obtenção dos dados de abundância relativa dos microrganismos da microbiota intestinal de pessoas que possuem obesidade ou diabetes tipo 2, e de pessoas saudáveis;
2) Pré-processamento, transformação e normalização dos dados obtidos;
3) Construção de modelos de Regressão Logística e *Random Forest* para predição do fenótipo e descoberta de biomarcadores;
4) Avaliação da performance preditiva e relevância dos biomarcadores encontrados.

3. METODOLOGIA

3.1. Obtenção dos dados

Os dados foram obtidos do website Kaggle, sendo que o conjunto utilizado foi o *Metagenomics*, oriundo do estudo de Pasolli et al. (2016). Esses dados são de livre acesso, e podem ser encontrados em: *https://www.kaggle.com/datasets/antaresnyc/metagenomics*. Os dados utilizados são da abundância relativa de microrganismos da microbiota intestinal de pessoas acometidas com determinadas doenças, contendo a abundância dos microrganismos em vários níveis da classificação taxonômica. Este conjunto de dados foi formado pela combinação de oito conjuntos distintos, onde seis vieram de estudos caso-controle que abordavam diferentes doenças, no caso, Cirrose, Doença Inflamatória Intestinal, Câncer do Colo Retal, Obesidade e Diabetes Tipo 2, sendo que dois estudos diferentes estudaram Diabetes. Os dados dos dois estudos restantes são apenas de pessoas saudáveis, e não são exclusivos da microbiota intestinal.

3.2. Separação dos dados de interesse

Neste trabalho, os dados de interesse eram dos estudos que abordaram a obesidade e diabetes tipo 2. Sendo assim, apenas os dados que eram sobre essas enfermidades foram separados dos demais. Os dados sobre obesidade são do estudo de LE CHATELIER et al. (2013), e os dados sobre diabetes tipo 2 são do estudo de QIN et al. (2012). Os dados foram separados utilizando a biblioteca DataFrames (v 1.3.4) da linguagem de programação Julia (v 1.7.3, Bezanson et al., 2012).

3.3. Pré-processamento dos dados

Aplicando a biblioteca DataFrames (v 1.3.4), foram renomeados os dados da coluna “disease” do conjunto de diabetes, para “n” nas amostras controle e para “disease” nas amostras dos casos. Como esses dados vieram de estudos diferentes, ambos usavam nomes distintos para classificar as amostras de diabéticas. Essa renomeação foi feita para padronizar os nomes nas colunas, garantindo que o algoritmo possa interpretar essas informações de maneira correta.

Ainda usando a mesma biblioteca, foi feito o mesmo processo de renomeação da coluna “disease” dos dados de obesidade para “n” nos controles e “disease” para os casos. As amostras que estavam marcadas como "leaness" também foram classificadas para “n”, a fim de diminuir a discrepância entre o número de amostras de cada classe. Foram separados os dados da abundância relativa dos metadados de ambos os conjuntos, e filtradas as colunas de metadados que não tinha informação nenhuma, retirando as colunas que possuíam linhas contendo “-” ou “nd”. Nos dados da abundância relativa, foram retiradas as colunas cuja abundância máxima era igual a zero, já que elas não seriam informativas, e também foram deixadas apenas as colunas que representavam a abundância no nível de gênero. Nesses dados, as abundâncias são cumulativas, ou seja, a abundância de uma classe taxonômica mais abrangente, como família, vai se diluindo entre as classes mais específicas, como gênero.

Os nomes das colunas seguem o padrão de possuir a primeira letra representando a classe, seguido do nome da classe a qual aquele organismo pertence. Como por exemplo, a coluna de nome: k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f Bacteroidaceae|g__Bacteroid es, representa a abundância relativa do gênero Bacteroides. Aplicando funções de processamento de texto da linguagem Julia (v 1.7.3, Bezanson et al., 2012), é possível remover separar o texto no caractere “|” e remover todos os textos que não começam com “g_”, deixando assim o nome da coluna apenas com o nome do gênero. Isso foi feito para melhorar a interpretabilidade dos dados, e diminuir a dimensionalidade. Esse procedimento foi realizado para ambos os conjuntos de dados.

Posteriormente, foi feita a separação dos dados de entrada, que são as abundâncias relativas de cada gênero, dos dados de saída, que é a classificação de cada amostra. Para isso, foi usada uma função da biblioteca MLJ (v 0.18.5, Blaom et al., 2020) que separou a coluna que continha os dados de saída “disease” das demais colunas. Ademais também foi feita a transformação da coluna saída, de texto para *OrderedFactor*. Essa transformação é necessária para que os modelos utilizados interpretem as saídas como categorias, e não apenas texto, e funcionem da maneira correta. Esse procedimento foi realizado para ambos os conjuntos de dados.

3.4. Normalização dos dados

Para que as técnicas utilizadas funcionem da maneira correta, é esperado que os dados de entrada sigam uma distribuição mais próxima da normal possível. Os dados originais não estavam normalizados, e isso poderia acarretar em um resultado enviesado e uma análise incorreta. Sendo assim, utilizando da biblioteca MLJ (v 0.18.5, Blaom et al., 2020), foi feita uma normalização por Z-Score, deixando os dados com média igual a zero e variância igual a um. Esse procedimento foi realizado em ambos os conjuntos de dados.

3.5. Separação dos conjuntos de treino e teste

Em um algoritmo de aprendizado de máquina, é necessário separar o conjunto de dados de interesse em conjuntos de treino e de teste, para que seja feita a validação correta do modelo construído, utilizando dados inéditos, que não fizeram parte do treinamento do modelo. Para isso, foi utilizada a biblioteca MLJ (v 0.18.5, Blaom et al., 2020), para separar ambos os conjuntos de dados, usando 70% dos conjuntos para treino e 30% para teste. A ordem das amostras também foi embaralhada aleatoriamente, garantindo que uma presença equivalente de ambas as classes no treino e no teste. Isso impede que o modelo acabe sendo treinado apenas nos dados de controle ou de caso, e não consiga generalizar para o problema inteiro.

3.6. Filtragem por correlação linear

A fim de melhorar a performance do modelo, foi feita filtragem de colunas que eram colineares com outras. Foi calculada a correlação linear de Pearson entre todas as colunas, usando a função *cor()*. Todas as colunas que possuíam correlação maior que 0.9 com outras foram separadas. Aquelas colunas que tinham essa correlação maior que 0.9 com uma ou mais colunas foram retiradas.

3.7. Construção do modelo de Regressão Logística e *Random Forest*

Usando da biblioteca MLJ (v 0.18.5, Blaom et al., 2020), foi construído o modelo utilizado de Regressão Logística. O modelo foi carregado usando os seguintes hiperparâmetros: *max_iter* = 3000, *penalty* = "l1", *solver* = "liblinear". Esses parâmetros foram escolhidos pois geraram os melhores resultados. A escolha do parâmetro *penalty* “l1” permite que seja feita uma seleção de atributos dentro do próprio modelo, já que com

essa penalidade é possível identificar quais foram os atributos mais relevantes para a predição. Essa funcionalidade foi usada para a descoberta de biomarcadores em ambos os conjuntos de dados.

Também foi construído o modelo de *Random Forest* usando a biblioteca MLJ (v 0.18.5, Blaom et al., 2020). Ele foi carregado com os parâmetros padrões.

Na Regressão Logística, o hiperparâmetro *C* do modelo foi o escolhido para ser ajustado usando o método de *GridSearch*. Esse hiperparâmetro é o limite de regularização que a penalidade "l1" usa, ele impõe um limite que a soma dos valores absolutos dos pesos ajustados não pode ser maior que o determinado valor. Isso garante que o modelo zere aquelas variáveis irrelevantes, na tentativa de deixar a soma menor que o limite, e previne o modelo de se ajustar muito ao ruído, e tenha um ajuste mais generalizado dos dados. O ajuste do hiperparâmetro possibilita que o modelo seja treinado e testado usando valores alterando entre 0.1 e 1.0, na escala logarítmica, organizados em um *Grid* de dimensão dez. Ou seja, cem valores para o hiperparâmetro *C* foram testados, por meio de *Stratified K-Fold Cross-Validation* no conjunto de treino, e o valor que gerou o menor número de classificações incorretas foi escolhido para ser usado no modelo final. A testagem por *Stratified K-Fold Cross-Validation*, separa o conjunto de treino em K partes, mantendo a porcentagem de cada classe em todas as partes, treina o modelo com o valor da vez em todas as K-1 partes, e testa na parte restante. Isso garante que o hiperparâmetro encontrado foi testado em todo o dado, e não apenas em uma parcela. Foram usadas cinco dobras nessa etapa. Esse ajuste foi feito para ambos os conjuntos de dados.

No modelo de *Random Forest*, o hiperparâmetro escolhido a ser ajustado foi o *n_estimators*, que representa o número de árvores que terá na floresta. Esse hiperparâmetro também foi ajustado por *GridSearch*, com um *Grid* de tamanho dez, usando *Stratified Cross-Validation* de cinco dobras. O número de árvores que gerou o maior valor de *auc*, ou seja, área sob a curva ROC, foi escolhido pelo algoritmo.

Após o ajuste dos hiperparâmetros de ambos os modelos, eles foram treinados nos dados de treino, por meio da biblioteca MLJ (v0.18.5, Blaom et al., 2020), usando *Cross-Validation* de cinco dobras, e o treino foi repetido 5 vezes. O mesmo procedimento foi realizado para ambos os conjuntos de dados.

3.8. Seleção de atributos e re-treinamento

Após o primeiro treino, aplicando a biblioteca DataFrames (v1.3.4), foi resgatado os coeficientes de cada variável, no caso da Regressão Logística, que representam a importância de cada uma delas para o ajuste do modelo, e os valores de importância de cada variável, no caso da *Random Forest*,. Foram identificadas aquelas variáveis cuja importância era igual a zero, ou seja, que foram irrelevantes para o ajuste do modelo final. Essas variáveis foram retiradas do conjunto de dados, deixando apenas aquelas variáveis que contribuíram para o ajuste do modelo.

Usando a biblioteca MLJ (v0.18.5, Blaom et al., 2020), foi refeito o treinamento de ambos os modelos com os dados contendo apenas as respectivas variáveis selecionadas, usando os mesmos parâmetros do primeiro treino. Esse procedimento foi realizado para ambos os conjuntos de dados.

3.9. Identificação de biomarcadores

Com o segundo treinamento realizado, na Regressão Logística, foram resgatados novamente os coeficientes das variáveis usadas. Aquelas variáveis que mais influenciaram positivamente ou negativamente no ajuste do modelo, foram investigadas.

No modelo de *Random Forest*, a identificação de biomarcadores foi feita por um método iterativo, onde 50% por cento das variáveis mais importantes eram usadas para treino, usando *Cross-Validation*. Esse procedimento foi realizado 5 vezes, cada vez usando apenas 50% das variáveis mais importantes, restando apenas 5 variáveis no final do processo. Essas 5 variáveis foram investigadas.

3.10. Validação e avaliação de performance

O modelo final construído foi validado usando o conjunto de testes que foi reservado, cujo os dados não fizeram parte da etapa do treinamento, ou seja, são dados inéditos para o modelo. Usando a biblioteca MLJ (v0.18.5, Blaom et al., 2020), foi feita a predição do fenótipo nesses dados inéditos.

O desempenho do modelo foi avaliado usando quatro métricas, *f1score,* acurácia,

precisão e sensitividade. A métrica *f1score* é a média harmônica entre precisão e sensitividade, a acurácia é a proporção de predições corretas do modelo, sensitividade mede dentre todos os casos positivos, o quanto o modelo acertou, e a precisão mede dentre todas as previsões positivas, quantas realmente eram positivas. Também foram geradas as matrizes de confusão, que mostram os valores de falso e verdadeiro, positivo e negativo, para ambos, e os gráficos de curva ROC com o valor da *auc*, que mede o quão bem o modelo consegue distinguir entre as classes.

3.11. Investigação de biomarcadores

Para todos os biomarcadores encontrados, foram gerados gráficos do tipo *boxplot*, usando a biblioteca StatsPlots (v.0.15.4), comparando a distribuição das abundâncias, entre casos e controles. Também foi realizado o teste estatístico de Mann-Whitney para averiguar se havia diferença significativa entre a abundância dos biomarcadores nos casos e nos controles.

4. RESULTADOS

4.1. Resultados de predição da Regressão Logística nos dados de obesidade

O modelo final foi treinado nos dados de treino de obesidade com as melhores variáveis selecionadas, e foi avaliado usando quatro diferentes métricas (Tabela 1). O modelo também teve o valor de *auc* e curva ROC (Figura 2), além da matriz de confusão gerada (Tabela 2).

Tabela 1: Resultados da validação do modelo no conjunto de testes nos dados de obesidade.

F1score	0.69
Acurácia	0.60
Sensitividade	0.88
Precisão	0.56
Área sob a curva ROC	0.59

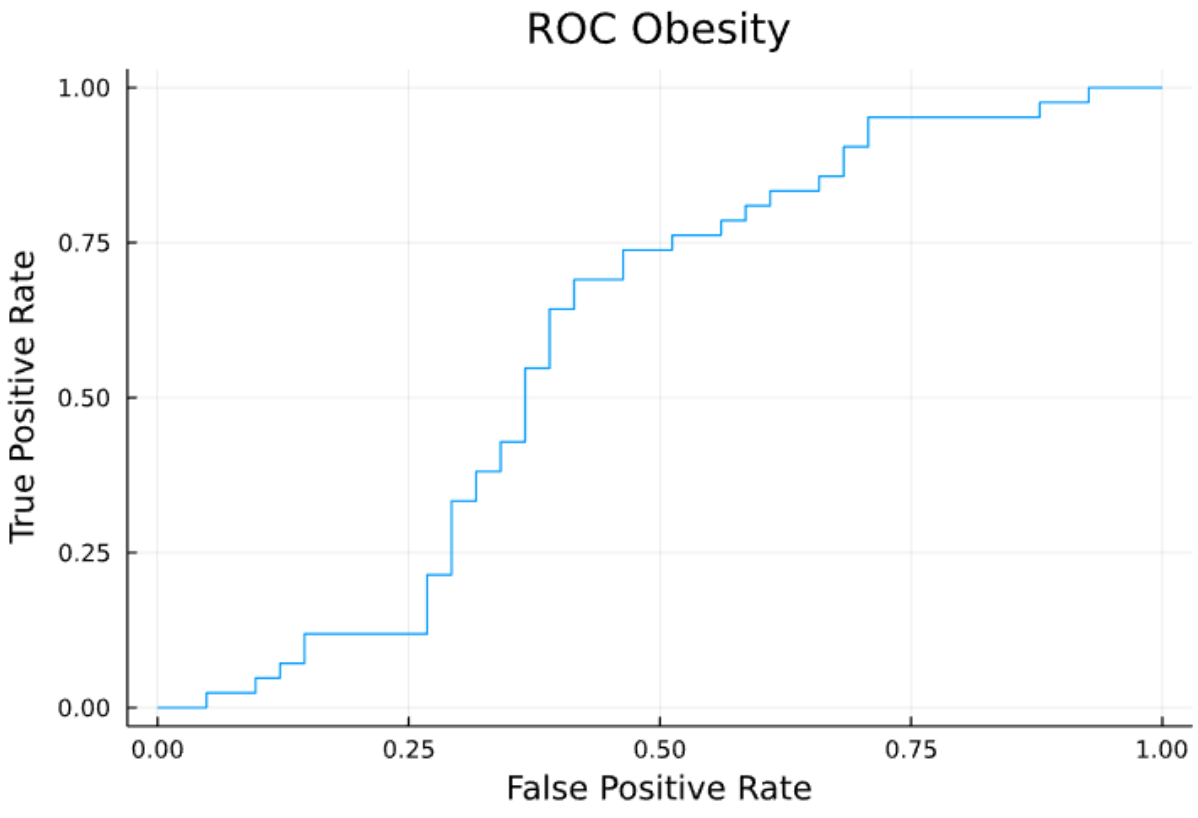
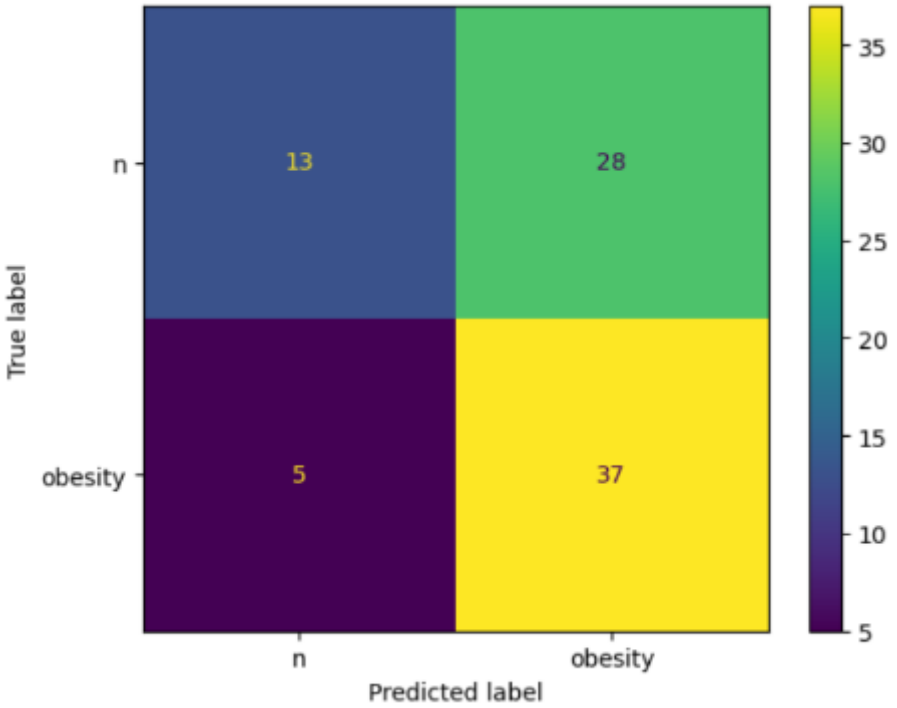


Figura 3: Curva ROC do modelo de Regressão Logística treinado nos dados de obesidade.

Tabela 2: Matriz de confusão do modelo de Regressão Logística treinado nos dados de obesidade.



4.2. Resultados de predição da Regressão Logística nos dados de diabetes

O modelo final foi treinado nos dados de treino de diabetes com as melhores variáveis selecionadas, e foi avaliado usando quatro diferentes métricas (Tabela 3). O modelo também teve o valor de *auc* e curva ROC (Figura 3), além da matriz de confusão gerada (Tabela 4).

Tabela 3: Resultados da validação do modelo no conjunto de testes nos dados de diabetes.

F1score	0.66
Acurácia	0.61
Sensitividade	0.75
Precisão	0.59
Área sob a curva ROC	0.70

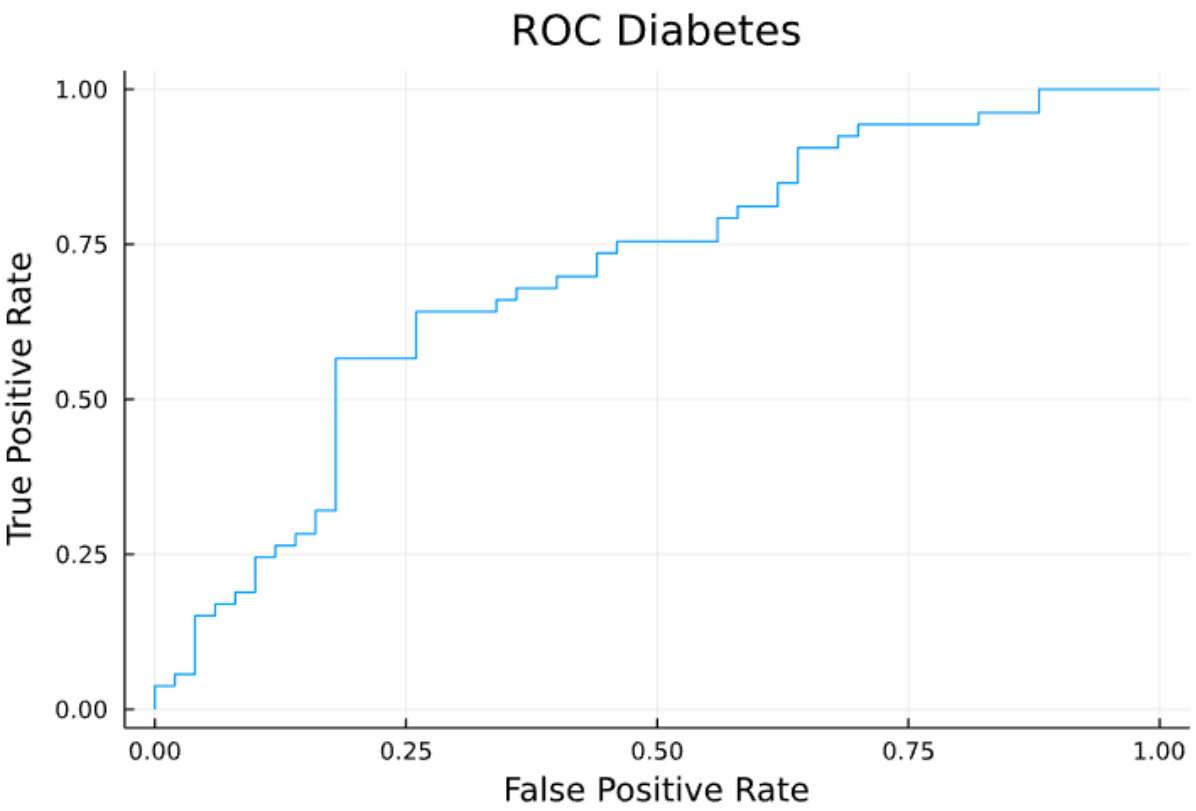
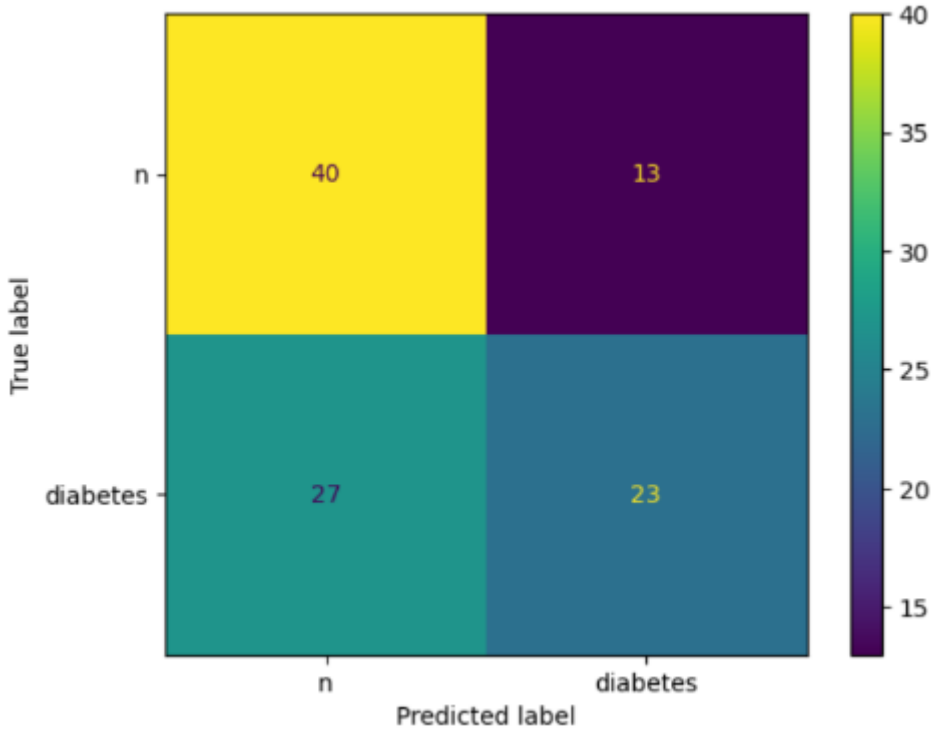


Figura 4: Curva ROC do modelo de Regressão Logística treinado nos dados de diabetes.

Tabela 4: Matriz de confusão do modelo de Regressão Logística treinado nos dados de diabetes.



4.3. Resultados de predição da Random Forest nos dados de obesidade

O modelo final foi treinado nos dados de treino de obesidade com as melhores variáveis selecionadas, e foi avaliado usando quatro diferentes métricas (Tabela 5). O modelo também teve o valor de *auc* e curva ROC (Figura 4), além da matriz de confusão gerada (Tabela 6).

Tabela 5: Resultados da validação do modelo no conjunto de testes nos dados de obesidade.

F1score	0.59
Acurácia	0.53
Sensitividade	0.65
Precisão	0.54
Área sob a curva ROC	0.56

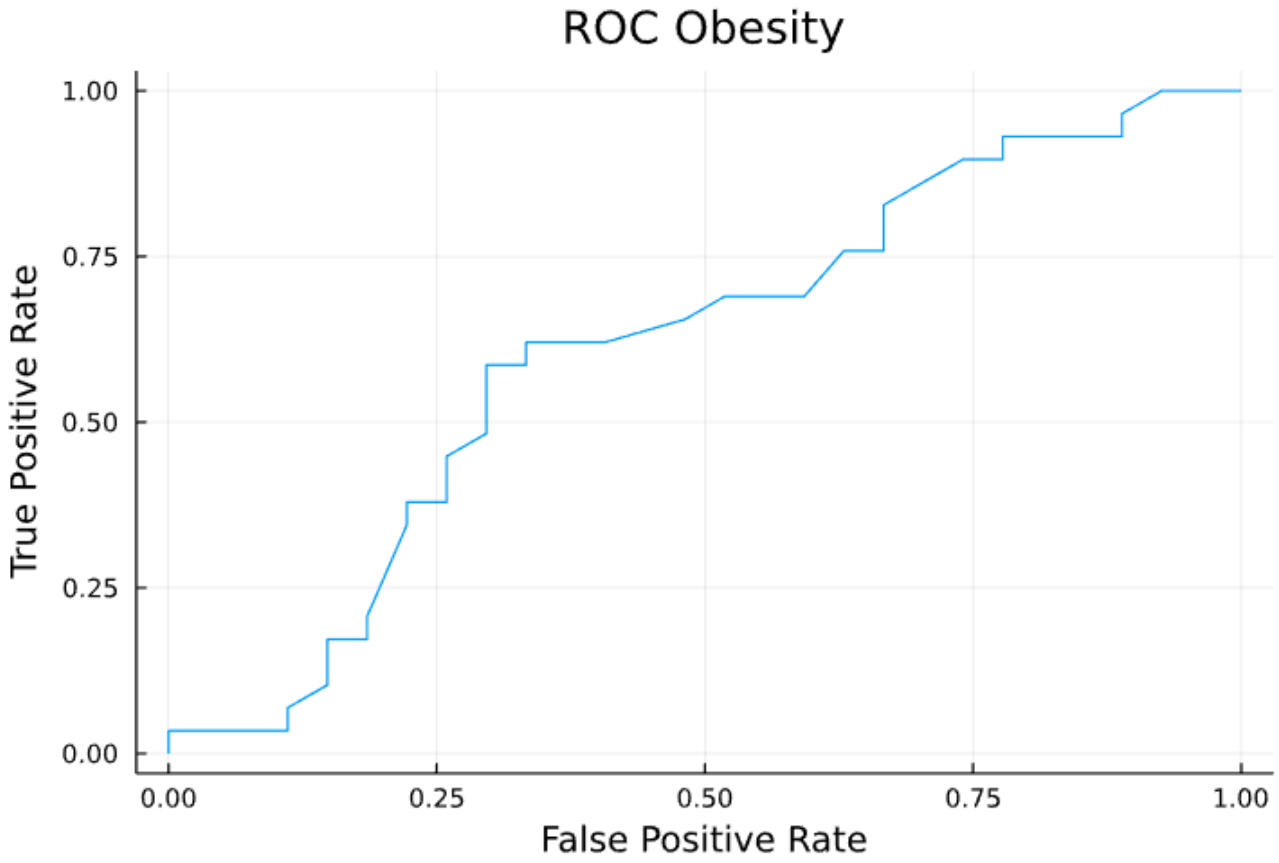
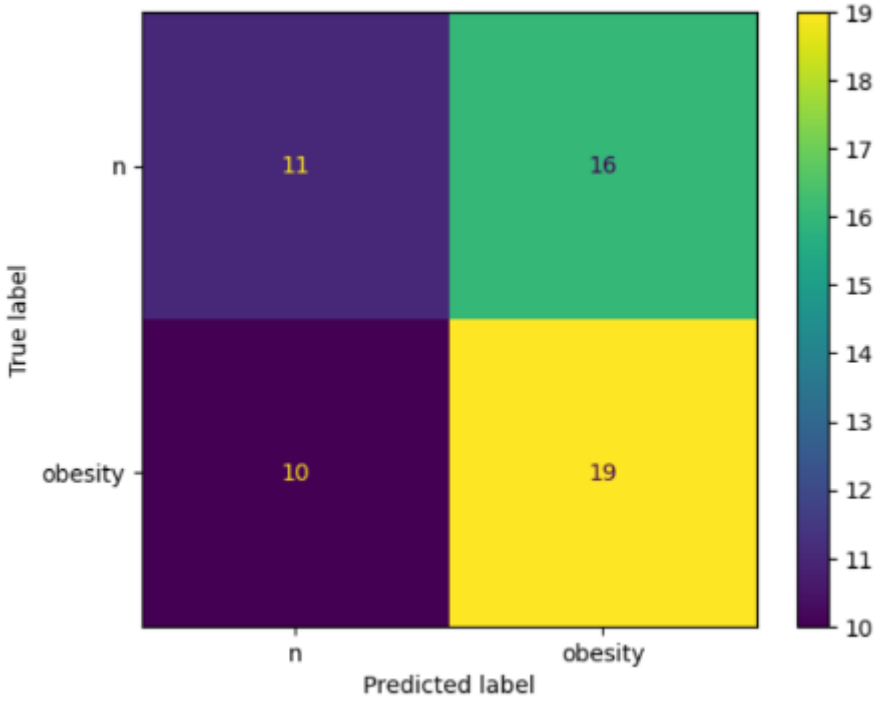


Figura 5: Curva ROC do modelo de *Random Forest* treinado nos dados de obesidade

Tabela 6: Matriz de confusão do modelo de *Random Forest* treinado nos dados de obesidade



-4.4. Resultados de predição da Random Forest nos dados de diabetes

O modelo final foi treinado nos dados de diabetes com as melhores variáveis selecionadas, e foi avaliado usando quatro diferentes métricas (Tabela 7). O modelo também teve o valor de *auc* e curva ROC (Figura 5), além da matriz de confusão gerada (Tabela 8).

Tabela 7: Resultados da validação do modelo no conjunto de testes nos dados de diabetes.

F1score	0.64
Acurácia	0.60
Sensitividade	0.58
Precisão	0.70
Área sob a curva ROC	0.73

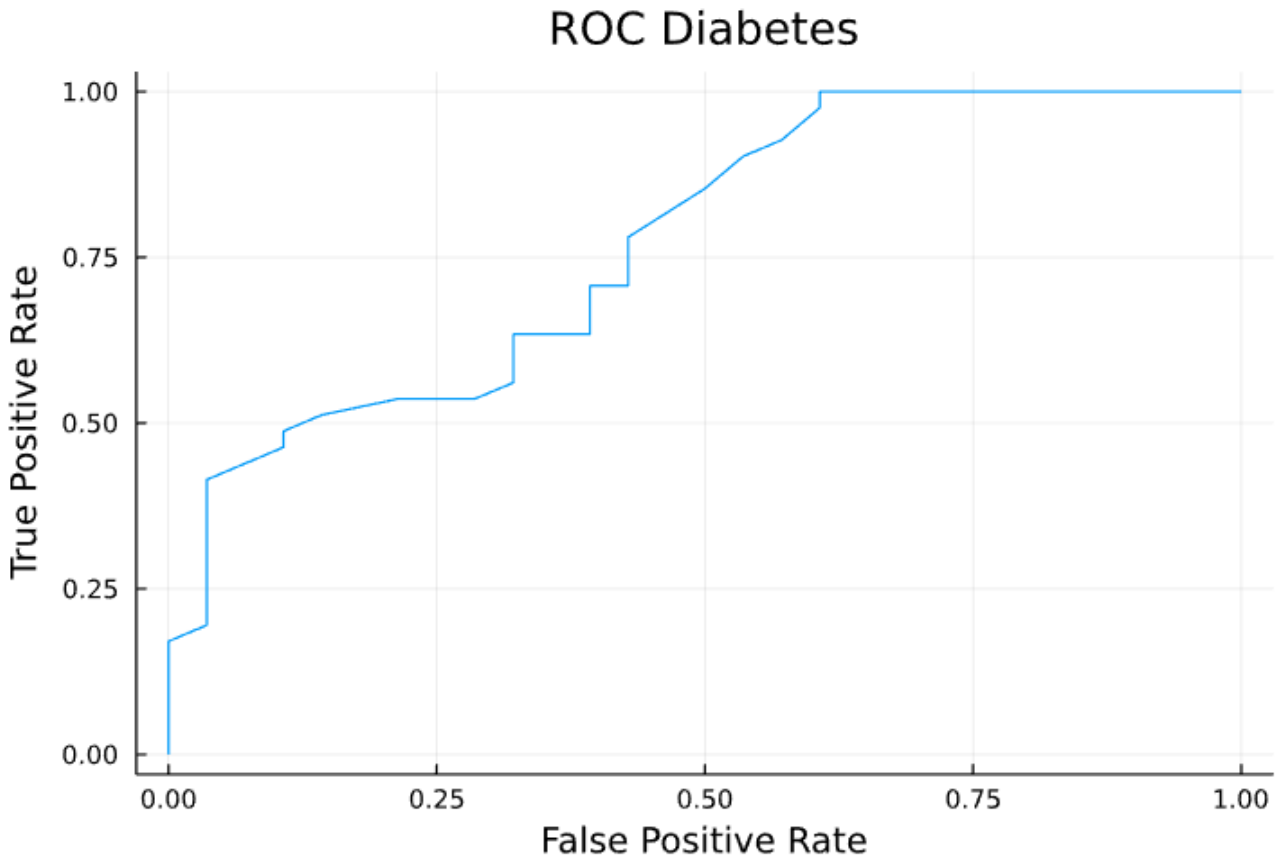
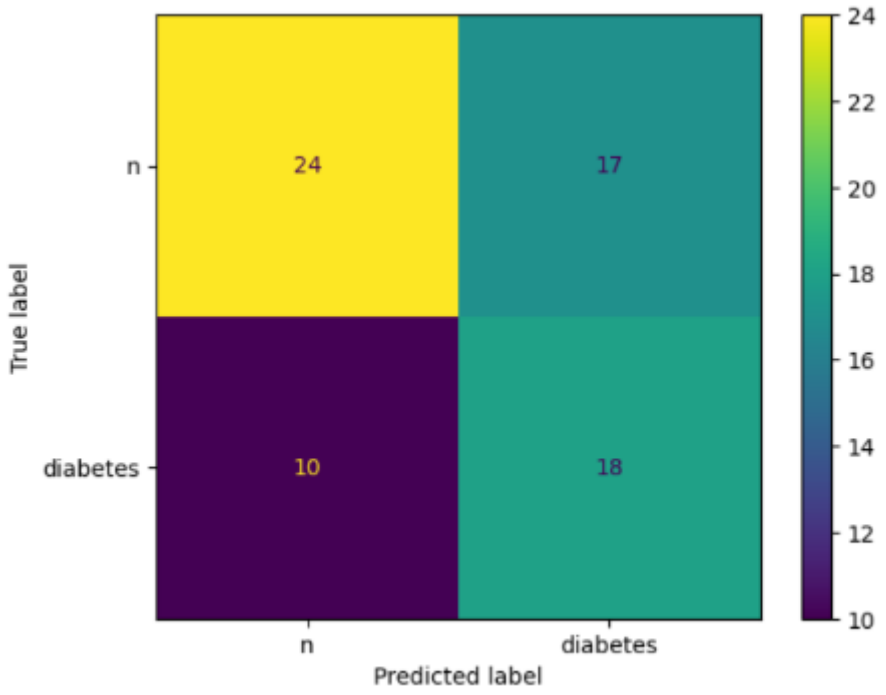


Figura 6: Curva ROC do modelo de *Random Forest* treinado nos dados de diabetes

Tabela 8: Matriz de confusão do modelo de *Random Forest* treinado nos dados de diabetes



4.5. Descoberta e investigação de biomarcadores utilizando Regressão Logística nos dados de obesidade

Por meio dos coeficientes das variáveis que o modelo de Regressão Logística retorna, podemos descobrir quais foram os gêneros que tiveram mais influência no ajuste do modelo final. Os gêneros *Deinococcus, Spiroplasma* e *Neisseria* tiveram os maiores coeficientes negativos, enquanto os gêneros *Porphyromonas, Bilophila* e *Ruminococcus* tiveram os maiores coeficientes positivos.

Os gêneros *Deinococcus* e *Porphyromonas* tiveram uma diferença significativa de abundância entre casos e controles, rejeitando a hipótese nula do teste de Mann-Whitney, com um intervalo de confiança de 95% (Tabela 9). Comparando os dois grupos, pode ser observada uma diferença na abundância do gênero *Deinococcus* (Figura 6) e *Porphyromonas* (Figura 7).

Tabela 9: P-valores dos biomarcadores encontrados usando Regressão Logística nos dados de obesidade.

Gênero	p-valor
Deinococcus	0,0003
Spiroplasma	0,0902
Neisseria	0,0902
Porphyromonas	0,0101
Bilophila	0,1397
Ruminococcus	0,1119

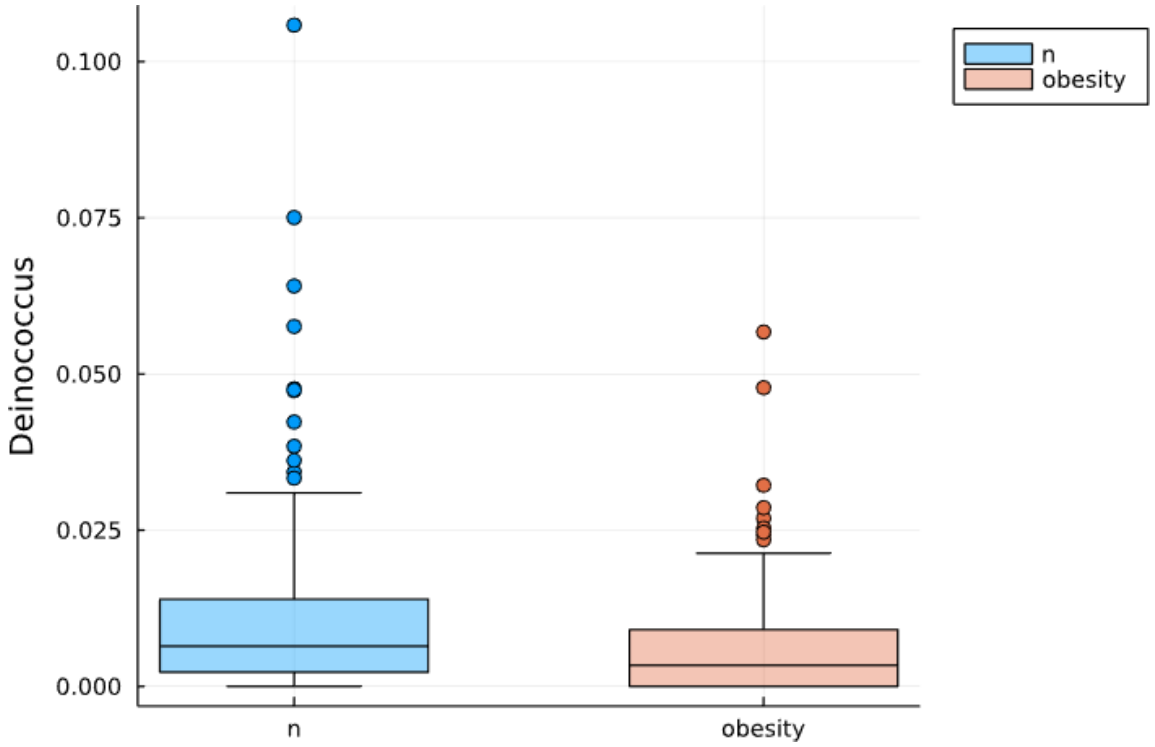


Figura 7: Distribuição da abundância do gênero *Deinococcus* nos dados de obesidade.

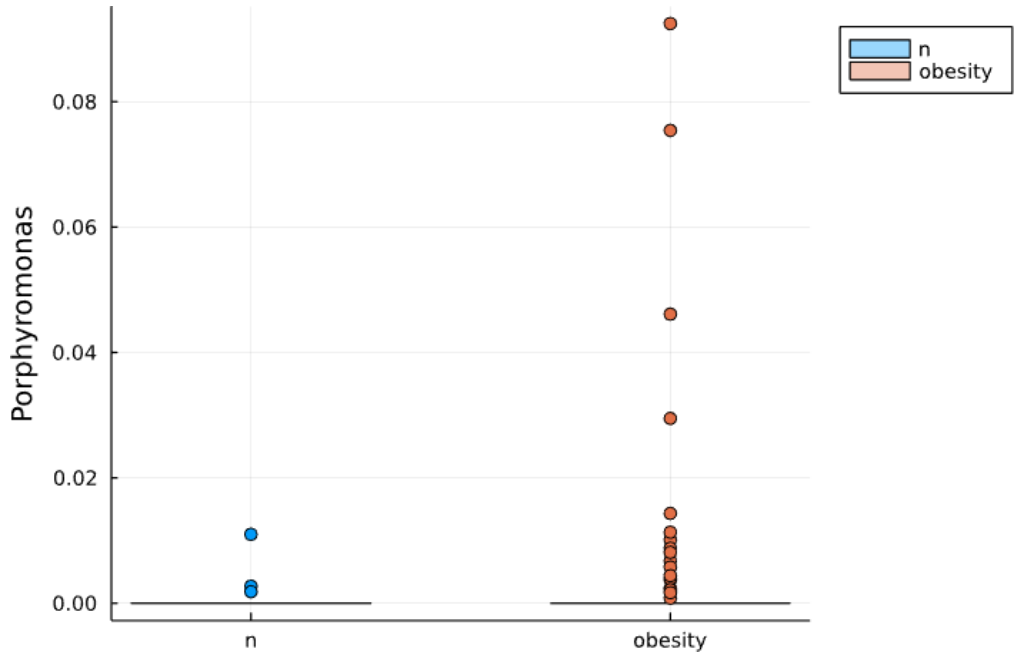


Figura 8: Distribuição da abundância do gênero *Porphyromonas* nos dados de obesidade.

4.6. Descoberta e investigação de biomarcadores utilizando Regressão Logística nos dados de diabetes

No caso da diabetes, também foram escolhidos os três mais influentes. Os gêneros *Bifidobacterium, Lactobacillus* e *Clostridium* tiveram os maiores coeficientes negativos, enquanto os gêneros *Flavonifractor, Peptostreptococcaceae_noname* e *Prevotella* os maiores coeficientes positivos. Vale mencionar que o quarto que mais influenciou negativamente foi o gênero *Bilophila*, que também apareceu nos mais influentes da obesidade, porém positivamente.

Os gêneros *Lactobacillus, Clostridium e Peptostreptococcaceae_noname* tiveram uma diferença significativa de abundância entre casos e controles, rejeitando a hipótese nula do teste de Mann-Whitney, com um intervalo de confiança de 95% (Tabela 10). Comparando os dois grupos, há uma abundância maior de *Lactobacillus* (Figura 8) e *Clostridium* (Figura 9) e menor de *Peptostreptococcaceae_noname* (Figura 10) no grupo dos casos.

Tabela 10: P-valores dos biomarcadores encontrados usando Regressão Logística nos dados de diabetes.

Gênero	p-valor
Bifidobacterium	0,1871
Lactobacillus	0,0004
Clostridium	0,0075
Flavonifractor	0,5595
Peptostreptococcaceae_noname	1,00E-04
Prevotella	0,06

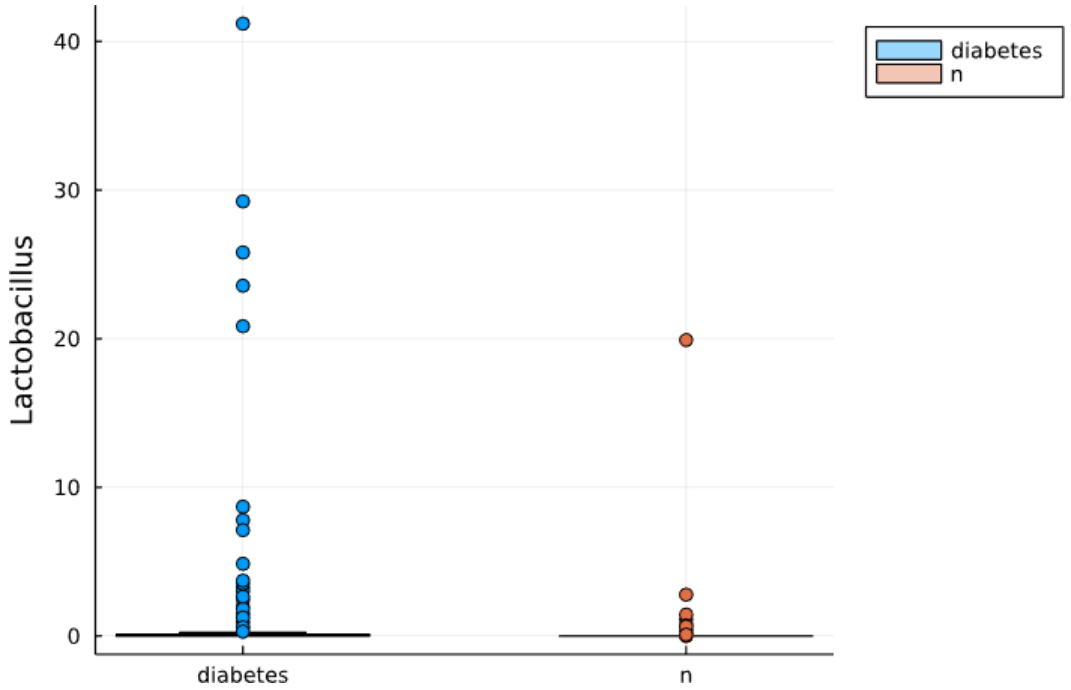


Figura 9: Distribuição da abundância de *Lactobacillus* nos dados de diabetes.

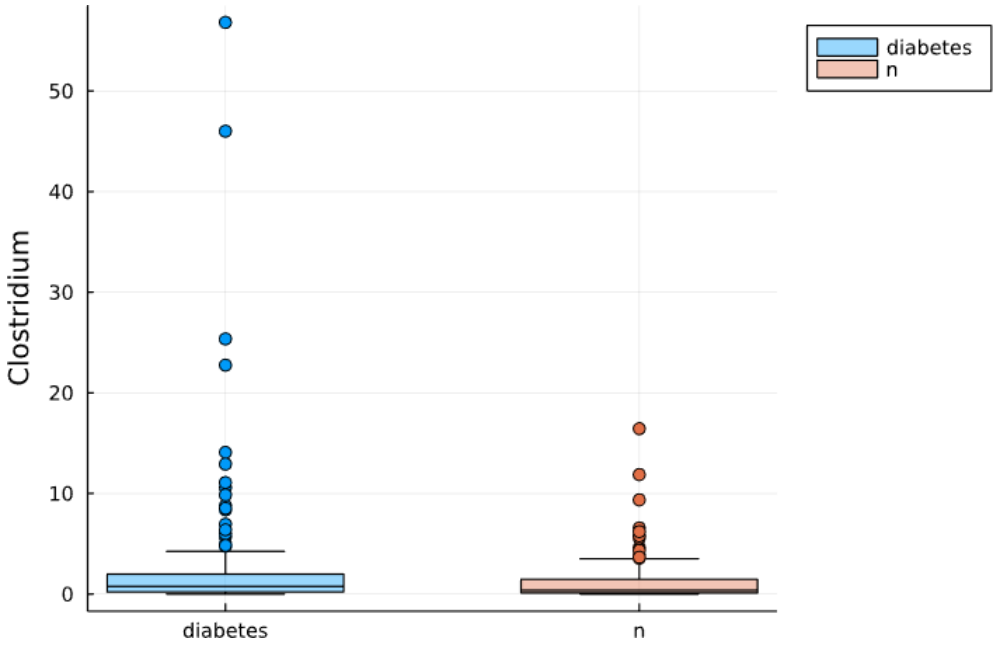


Figura 10: Distribuição da abundância do gênero *Clostridium* nos dados de diabetes

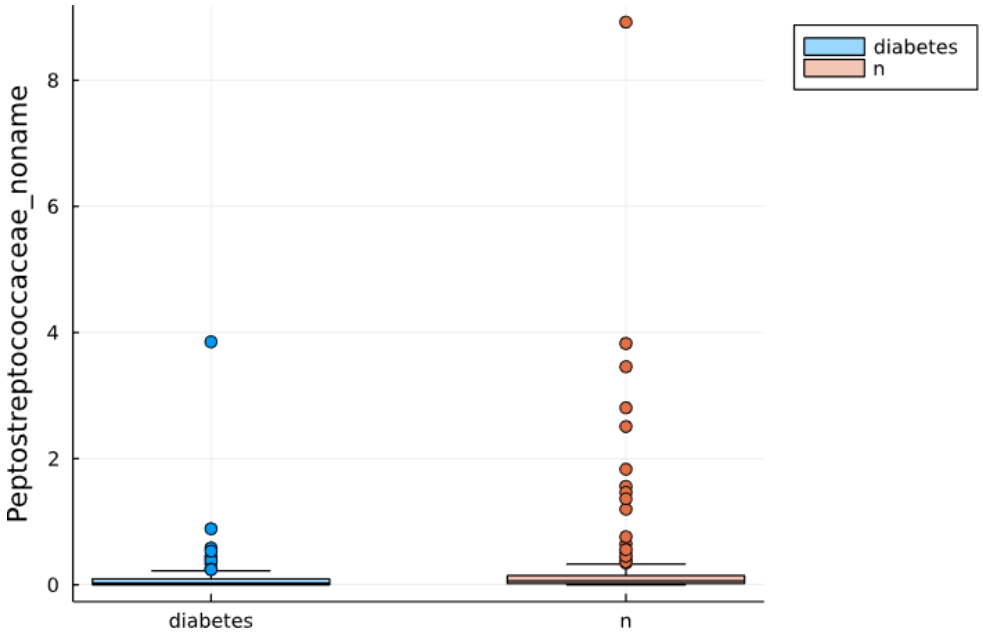


Figura 11: Distribuição da abundância do gênero *Peptostreptococcaceae_noname* nos dados de diabetes.

4.7. Descoberta e investigação de biomarcadores utilizando Random Forest nos dados de obesidade

Por meio do método iterativo realizado, foi possível identificar cinco gêneros que tiveram grande importância no ajuste do modelo final. Os gêneros mais importantes foram *Butyrivibrio, Veillonella, Anaerotruncus, Haemophilus, Subdoligranulum.*

Pelo teste de Mann-Whitney, todos os gêneros tiveram uma diferença significativa, rejeitando a hipótese nula, com um intervalo de confiança de 95% (Tabela 11). As distribuições estão exibidas nas figuras abaixo.

Tabela 11: P-valores dos biomarcadores encontrados usando *Random Forest* nos dados de obesidade.

Gênero	p-valor
Butyrivibrio	0,0157
Veillonella	0,0114
Anaerotruncus	0,0134
Haemophilus	0,0123
Subdoligranulum	0,0325

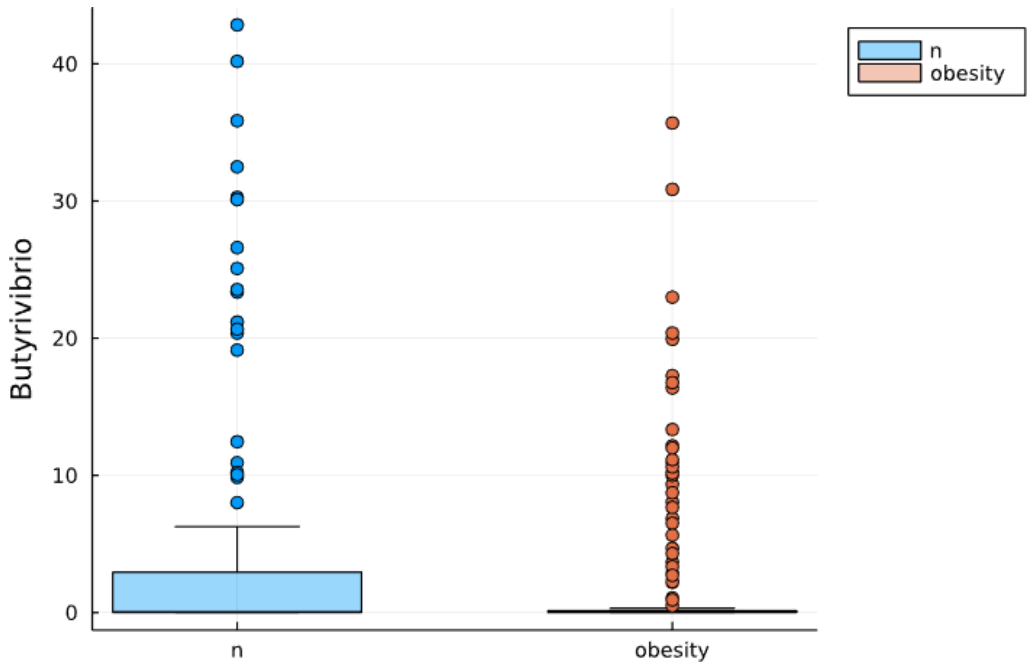


Figura 12: Distribuição da abundância do gênero *Butyrivibrio* nos dados de obesidade.

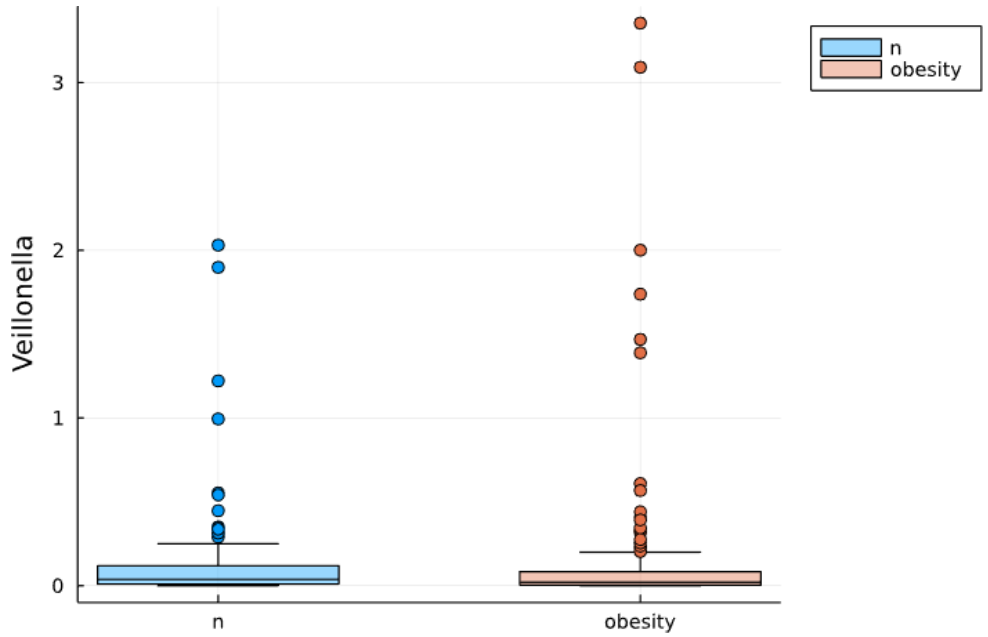


Figura 13: Distribuição da abundância do gênero *Veilonella* nos dados de obesidade.

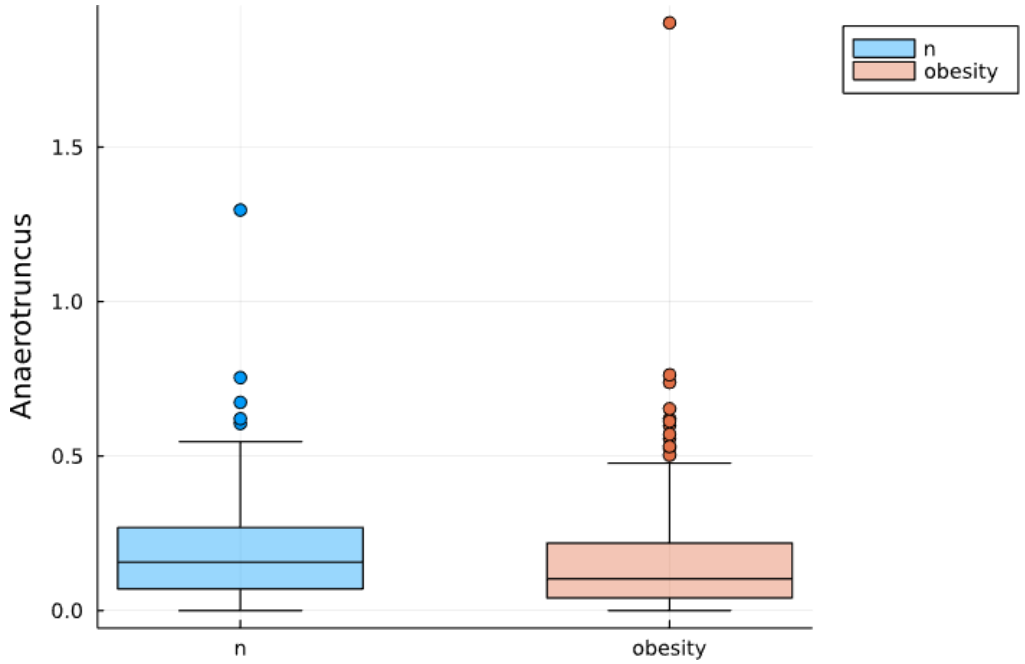


Figura 14: Distribuição da abundância do gênero *Anaerotruncus* nos dados de obesidade.

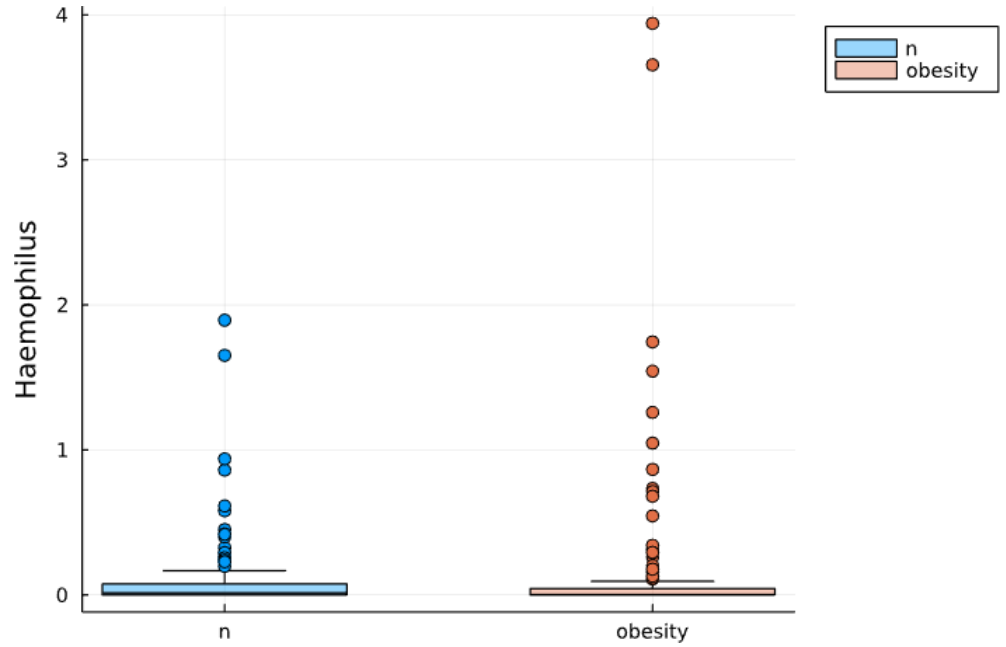


Figura 15: Distribuição da abundância do gênero *Haemophilus* nos dados de obesidade.

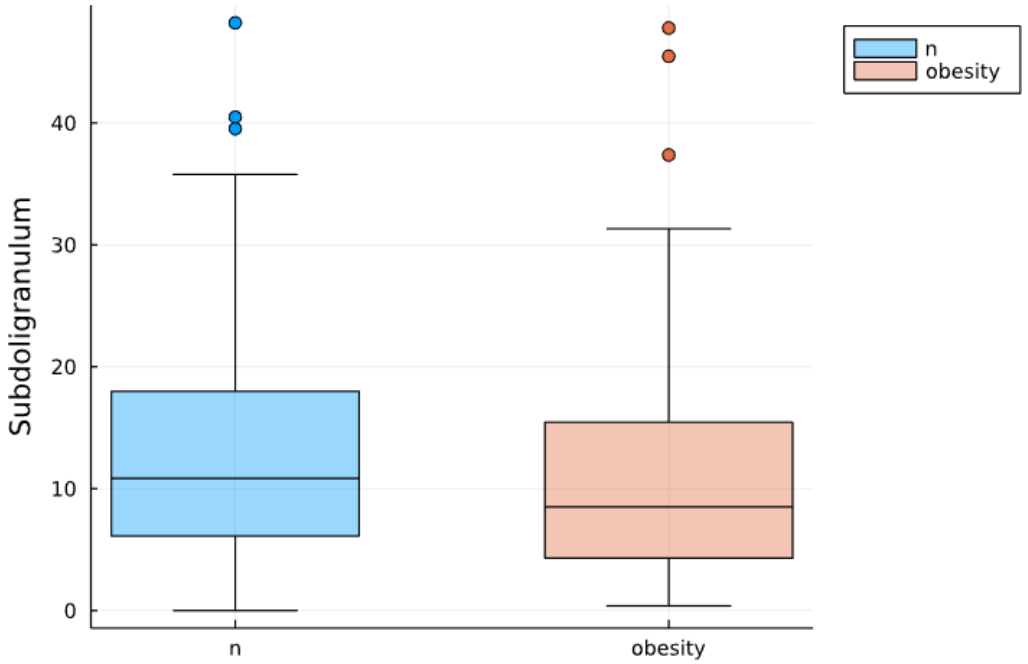


Figura 16: Distribuição da abundância do gênero *Subdoligranulum* nos dados de obesidade.

4.8. Descoberta e investigação de biomarcadores utilizando Random Forest nos dados de diabetes

Por meio do método iterativo realizado, foi possível identificar cinco gêneros que tiveram grande importância no ajuste do modelo final. Os gêneros mais importantes foram *Subdoligranulum, Bilophila, Parabacteroides, Roseburia e Peptostreptococcaceae_noname.*

Pelo teste de Mann-Whitney, apenas os gêneros *Parabacteroides, Roseburia e Peptostreptococcaceae_noname* tiveram uma diferença significativa, rejeitando a hipótese nula, com um intervalo de confiança de 95% (Tabela 12). O gênero *Parabacteroides* está mais abundante nos casos de microbiota derivadas de pessoas com diabetes (Figura 16), enquanto os gêneros *Roseburia* (Figura 17) e *Peptostreptococcaceae_noname* (Figura 18) estão mais abundantes nos controles.

Tabela 12: P-valores dos biomarcadores encontrados usando *Random Forest* nos dados de diabetes.

Gênero	p-valor
Subdoligranulum	0,2034
Bilophila	0,1023
Parabacteroides	0,0001

Roseburia	1,00E-05
Peptostreptococcaceae_noname	1,00E-04

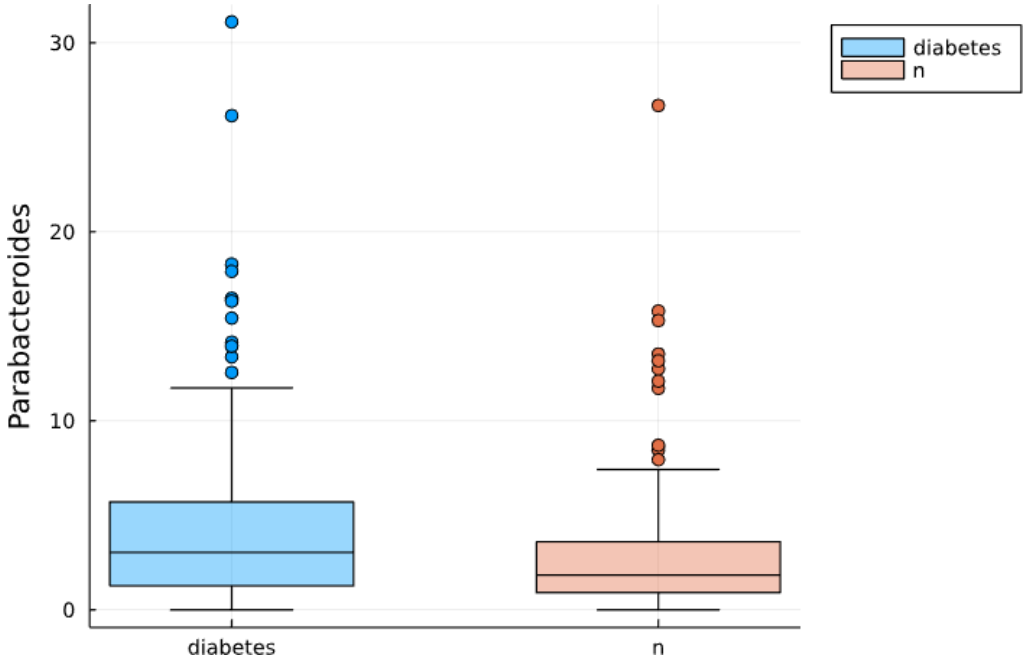


Figura 16: Distribuição da abundância do gênero *Parabacteroides* nos dados de diabetes.

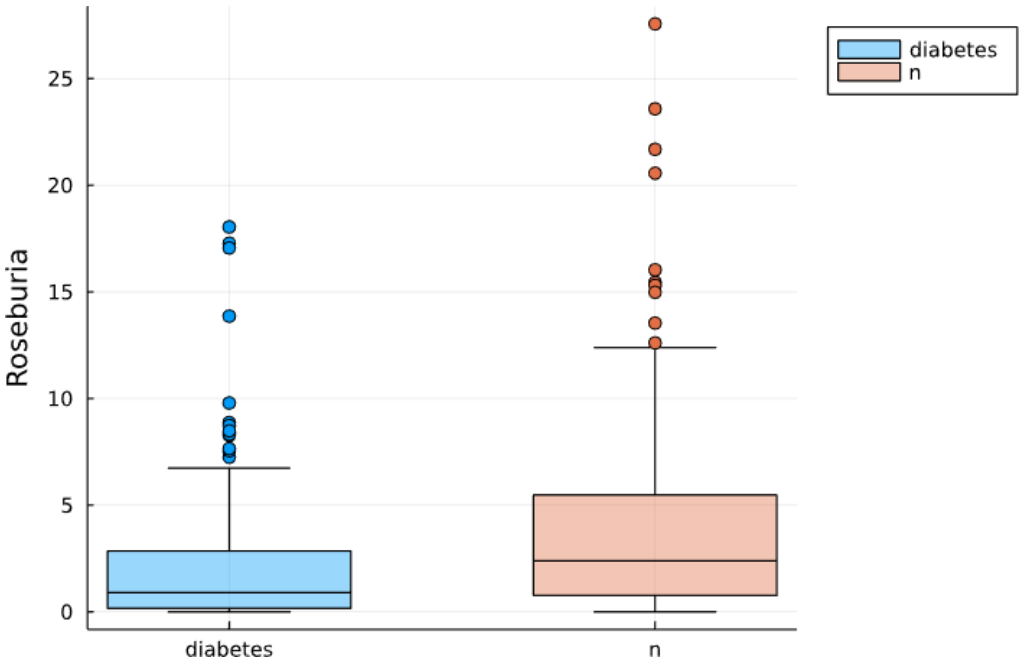


Figura 17: Distribuição da abundância do gênero *Roseburia* nos dados de diabetes.

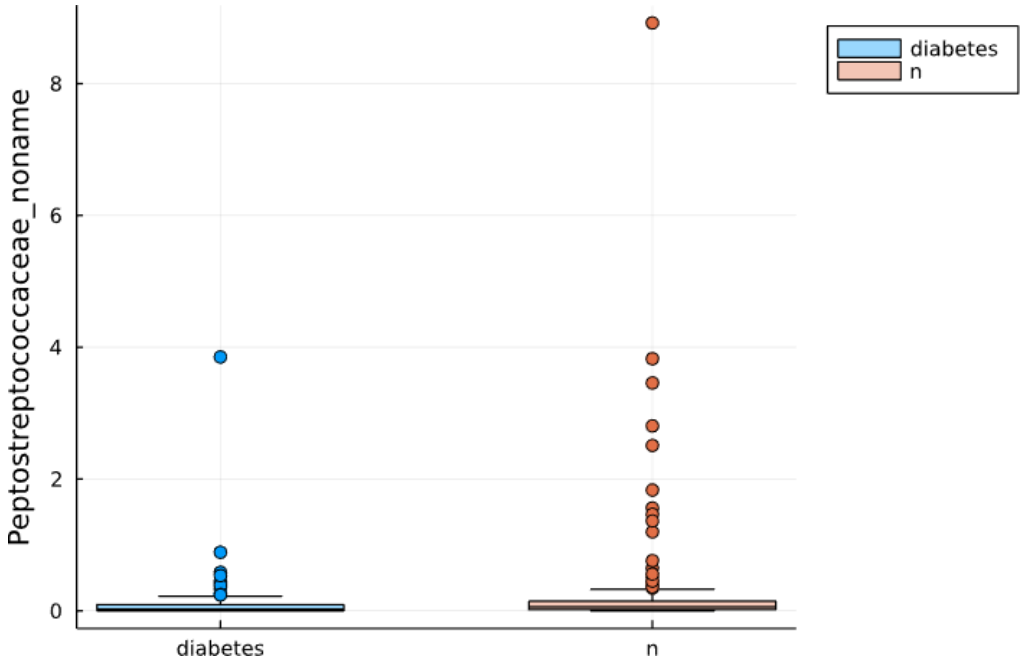


Figura 18: Distribuição da abundância do gênero *Peptostreptococcaceae_noname* nos dados de diabetes.

5. DISCUSSÃO

Com o aumento na geração de dados sobre a microbiota intestinal, tanto de pessoas saudáveis, quanto de pessoas com diferentes fenótipos, o uso de métodos de aprendizado de máquina pode vir a trazer novas descobertas sobre os mecanismos da microbiota intestinal em uma doença, possibilitando a predição da doença, ou de fatores associados, e da descoberta de biomarcadores associados a uma condição (MARCOS-ZAMBRANO et al., 2021).

Porém, existem ainda muitas limitações, principalmente quando falamos na diferença da microbiota intestinal das pessoas. Diversos fatores determinam a composição e estrutura da microbiota intestinal, como dieta, região do mundo, fatores genéticos, idade e sexo (LOZUPONE et al., 2012). Isso dificulta a construção de um modelo bem generalizado, já que muitas vezes esses fatores determinantes podem agir como fatores de desordem para o modelo, tendo mais poder na discriminação das categorias do que as variáveis alvo, que no caso eram as abundâncias relativas dos genêros (WIRBEL et al., 2021).

Outro fator limitante se dá nas diferentes técnicas usadas em cada estudo. A maioria dos estudos usam o gene 16S do RNA ribossomal das células procarióticas para identificação dos gêneros bacterianos. Porém, na maioria dos casos, apenas alguns fragmentos do gene 16S rRNA são sequenciados (maioritariamente as regiões hipervariáveis V3 e V4), o que limita a identificação dos géneros (Figura 19). Também, quais regiões do gene 16S são sequenciados diferem entre os estudos. Outros estudos, ainda, sequenciam o genoma inteiro da bactéria, por meio de sequenciamento *shotgun* dos metagenomas, como é o caso do estudo de Qin et al. (2012). Essas diferentes técnicas acabam refletindo na resolução e qualidade do dado usado, que pode gerar inconsistência nos resultados.

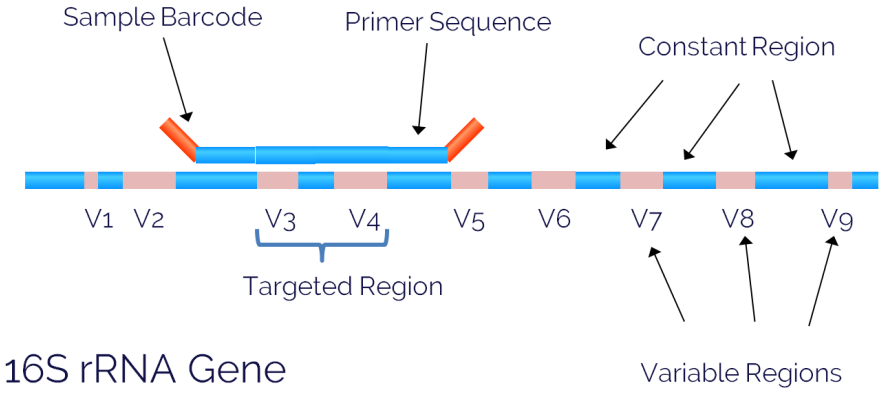


Figura 19: Ilustração da região mais comumente sequenciada do gene 16S. Fonte: CD Genomics, USA. 2021

Apesar dessas dificuldades, ambos os modelos conseguiram aprender a diferenciar os casos dos controles, sendo que os modelos de diabetes tiveram um desempenho significativamente melhor do que os de obesidade. Comparando os resultados dos modelos de obesidade deste trabalho, com os do estudo de Pasolli et al. (2016), que usou o mesmo conjunto de dados, temos números bem próximos, apesar da construção do modelo de *Random Forest* do estudo citado ter sido um pouco diferente. Sze e Schloss. (2016) realizaram uma meta-análise de diversos estudos de microbiota intestinal de indivíduos obesos, e perceberam que fatores associados como diferença na diversidade *alpha* na proporção de *Firmicutes* e *Bacteroidetes* ou riqueza da microbiota tinham sim uma diferença entre obesos e não-obesos, mas ela era muito pequena, e essa diferença era inconsistente entre os estudos. Eles também construíram um modelo de *Random Forest*, e conseguiram resultados semelhantes ao deste trabalho, mesmo usando maiores conjuntos de dados. Como citado anteriormente, as diferenças de técnicas e população estudadas limitam a generalização e descoberta de um consenso sobre as diferenças entre a microbiota de indivíduos obesos e não obesos.

Ainda comparando com o estudo de Pasolli et al. (2016), considerando os dados de diabetes, os resultados também foram similares. Os melhores números de desempenho deste conjunto podem ser associados ao tipo de sequenciamento usado, que foi do tipo *shotgun*, que gera maior cobertura do DNA do microrganismo, resultando em uma maior resolução, além do tamanho do conjunto de dados que era consideravelmente maior que o da obesidade.

Tanto a Regressão Logística quanto a *Random Forest* tiveram resultados próximos nas tarefas de predição, com os modelos de obesidade tendo um desempenho inferior em ambos. A Regressão é um modelo infinitamente mais simples que a *Random Forest*, porém sofre com multicolinearidade das variáveis. Já a *Random Forest* pode levar ao *overfit* mais facilmente, devido ao algoritmo ganancioso de construção das árvores, que sempre busca pelo melhor resultado. Ademais, a limitação está muito mais no conjunto de dados do que nos modelos, em relação às tarefas de predição.

Ambos os modelos conseguiram encontrar biomarcadores para as duas doenças, sendo que treze biomarcadores encontrados têm uma diferença significativa de abundância entre as duas classes. Nos biomarcadores encontrados pela *Random Forest*, destaca-se o gênero *Subdoligranulum*, que é benéfico e mais abundante nos saudáveis, e apareceu tanto na lista da

obesidade quanto na da diabetes, apontando uma possível relação entre as doenças. Além disso, o gênero *Bilophila*, que é maléfico e mais abundante nos indivíduos obesos, apareceu na lista da obesidade encontrada pela Regressão Logística, e na lista da diabetes encontrada pela *Random Forest*, apesar de não ter diferença significativa neste caso.

Microrganismos do gênero *Subdoligranulum* são benéficos ao corpo humano, tendo um efeito probiótico no metabolismo do hospedeiro. Também foi observada uma correlação positiva entre a presença dessas bactérias e a riqueza microbiana no trato intestinal, além de menor taxa de gordura visceral em indivíduos com alta abundância de *Subdoligranulum*. Porém, testes *in vivo* não obtiveram resultados positivos, quando observada a diferença de peso corporal, gordura, metabolismo de glucose e de lipídios, em ratos que foram administrados *Subdoligranulum variabile* por oito semanas. Esses resultados podem ser em decorrência da espécie e linhagem usada. Tecnologias de sequenciamento com boa resolução a nível de espécie e um maior entendimento do papel e atuação de microrganismos específicos na microbiota intestinal são avanços necessários para o desenvolvimento do conhecimento nessa área (VAN HUL et al., 2020)

Uma espécie do gênero *Bilophila* tem sinergia com uma dieta de altas calorias, e promove a inflamação, disfunção da barreira intestinal e alteração no metabolismo de ácido biliar. A administração do probiótico *Lactobacillus rhamnosus CNCM I-3690* limitou a ação da *Bilophila wadsworthia*, restituindo a barreira intestinal e reduzindo a inflamação, reforçando o impacto negativo dessa espécie na saúde humana (NATIVIDAD et al., 2018).

Não houve consenso significativo nos biomarcadores encontrados entre os dois modelos, contudo, a investigação dos gêneros que possuíam uma diferença significativa entre os dois grupos, pode vir a desvendar novos mecanismos e auxiliar no melhor entendimento das condições estudadas.

6. CONCLUSÕES

O uso de algoritmos de aprendizado de máquina em dados de abundância relativa, oriundos de estudos de caso controle sobre obesidade e diabetes tipo 2, obteve resultados semelhantes a outros estudos na tarefa de predição. Apesar das limitações do tamanho do conjunto de dados, diferenças nos métodos entre os estudos, e o pouco conhecimento sobre a microbiota intestinal e seu papel, há um potencial real do uso desses algoritmos na prevenção, diagnóstico e tratamento dessas doenças.

Dentre todos os biomarcadores encontrados, dois apareceram duas vezes, e contém estudos sobre o papel dos mesmos na obesidade. Mesmo com a limitação no âmbito dos dados, os algoritmos tiveram sucesso em identificar biomarcadores relevantes.

Por fim, a linguagem utilizada se mostrou eficiente e suficiente para a realização deste trabalho, mesmo sendo uma linguagem extremamente nova, e que ainda vem sendo atualizada constantemente.

7. REFERÊNCIAS

SEKIROV, I. et al. Gut Microbiota in Health and Disease. **Physiological Reviews**, v. 90, n. 3, p. 859–904, jul. 2010.

LEY, R. E. et al. Human gut microbes associated with obesity. **Nature**, v. 444, n. 7122, p. 1022–1023, dez. 2006.

QIN, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. **Nature**, v. 490, n. 7418, p. 55–60, 26 set. 2012.

ALVES, L. DE F. et al. Metagenomic Approaches for Understanding New Concepts in Microbial Science. **International Journal of Genomics**, v. 2018, p. 1–15, 23 ago. 2018.

WHEELER, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. **Nature**, v. 452, n. 7189, p. 872–876, abr. 2008.

KCHOUK, M.; GIBRAT, J. F.; ELLOUMI, M. Generations of Sequencing Technologies: From First to Next Generation. **Biology and Medicine**, v. 09, n. 03, 2017.

ROUMPEKA, D. D. et al. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. **Frontiers in Genetics**, v. 8, 6 mar. 2017.

SHARPTON, T. J. An introduction to the analysis of shotgun metagenomic data. **Frontiers in Plant Science**, v. 5, 16 jun. 2014.

LOZUPONE, C. A. et al. Diversity, stability and resilience of the human gut microbiota. **Nature**, v. 489, n. 7415, p. 220–230, set. 2012.

CLAESSON, M. J. et al. Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. **PLoS ONE**, v. 4, n. 8, p. e6669, 20 ago. 2009.

ZUO, T. et al. Alterations in Gut Microbiota of Patients With COVID-19 During Time of Hospitalization. **Gastroenterology**, v. 159, n. 3, p. 944- 955.e8, 1 set. 2020.

KARTAL, E. et al. A faecal microbiota signature with high specificity for pancreatic cancer. **Gut**, p. gutjnl-2021-324755, 8 mar. 2022.

WILLING, B. P. et al. A Pyrosequencing Study in Twins Shows That Gastrointestinal Microbial Profiles Vary With Inflammatory Bowel Disease Phenotypes. **Gastroenterology**, v. 139, n. 6, p. 1844-1854.e1, dez. 2010.

CHANG, J. et al. Decreased Diversity of the Fecal Microbiome in RecurrentClostridium difficile–Associated Diarrhea. **The Journal of Infectious Diseases**, v. 197, n. 3, p. 435–438, fev. 2008.

Obesity and overweight. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>. Acesso em: 30 nov. 2022.

TAGLIABUE, A.; ELLI, M. The role of gut microbiota in human obesity: Recent findings and future perspectives. **Nutrition, Metabolism and Cardiovascular Diseases**, v. 23, n. 3, p. 160–168, 1 mar. 2013.

Diabetes. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/diabetes>. Acesso em: 29 nov. 2022.

CHATTERJEE, A.; GERDES, M. W.; MARTINEZ, S. G. Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. **Sensors**, v. 20, n. 9, p. 2734, jan. 2020.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 44, n. 1.2, p. 206–226, jan. 2000.

MAHESH, B. Machine Learning Algorithms - A Review. **International Journal of Science and Research,** v. 9, n. 1, p. 6, 2018.

AKINSOLA, J. E. T. Supervised Machine Learning Algorithms: Classification and

Comparison. **International Journal of Computer Trends and Technology (IJCTT)**, v. 48, p. 128–138, 8 jun. 2017.

DHAR, S. et al. **On-Device Machine Learning: An Algorithms and Learning Theory Perspective**. arXiv, , 24 jul. 2020. Disponível em: <http://arxiv.org/abs/1911.00623>. Acesso em: 26 nov. 2022

SINGH, A.; THAKUR, N.; SHARMA, A. **A review of supervised machine learning algorithms**. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). **Anais**... Em: 2016 3RD INTERNATIONAL CONFERENCE ON COMPUTING FOR SUSTAINABLE GLOBAL DEVELOPMENT (INDIACOM). mar. 2016.

ABU-MOSTAFA, Y. S. **Learning from Data. A Short Course**. S.l.: AMLbook, 2012.

KLEINBAUM, D. G.; KLEIN, M. **Logistic Regression**. New York, NY: Springer New York, 2010.

RANSTAM, J.; COOK, J. A. LASSO regression. **British Journal of Surgery**, v. 105, n. 10, p. 1348, 1 set. 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Random Forests. Em: HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. (Eds.). **The Elements of Statistical Learning**. Springer Series in Statistics. New York, NY: Springer New York, 2009. p. 587–604.

HUTTENHOWER, C. et al. Structure, function and diversity of the healthy human microbiome. **Nature**, v. 486, n. 7402, p. 207–214, jun. 2012.

MARCOS-ZAMBRANO, L. J. et al. Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. **Frontiers in Microbiology**, v. 12, 2021.

STATNIKOV, A. et al. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. **Bioinformatics**, v. 21, n. 5, p. 631–643, 16

set. 2004.

PASOLLI, E. et al. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. **PLOS Computational Biology**, v. 12, n. 7, p. e1004977, 11 jul. 2016.

BEZANSON, J. et al. **Julia: A Fast Dynamic Language for Technical Computing**. arXiv, , 23 set. 2012. Disponível em: <http://arxiv.org/abs/1209.5145>. Acesso em: 27 nov. 2022

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

BLAOM, A. et al. MLJ: A Julia package for composable machine learning. **Journal of Open Source Software**, v. 5, n. 55, p. 2704, 7 nov. 2020.

WIRBEL, J. et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. **Genome Biology**, v. 22, n. 1, p. 93, 30 mar. 2021.

SZE, M. A.; SCHLOSS, P. D. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. **mBio**, v. 7, n. 4, p. e01018-16, 23 ago. 2016.

VAN HUL, M. et al. From correlation to causality: the case of *Subdoligranulum*. **Gut Microbes**, v. 12, n. 1, p. 1849998, 9 nov. 2020.

NATIVIDAD, J. M. et al. Bilophila wadsworthia aggravates high fat diet induced metabolic dysfunctions in mice. **Nature Communications**, v. 9, n. 1, p. 2802, 18 jul. 2018.

APÊNDICE A – Tabela dos coeficientes das variáveis do modelo de Regressão Logística

OBESIDADE

Genus	Values
Deinococcus	-0,742009
Spiroplasma	-0,447768
Neisseria	-0,351575
Ruminococcus	0,427802
Bilophila	0,570838
Porphyromonas	0,859379

DIABETES

Genus	Values
Bifidobacterium	-0,707830
Lactobacillus	-0,511474
Clostridium	-0,395307
Prevotella	0,344502
Peptostreptococcaceae_noname	0,636095
Flavonifractor	0,752718

APÊNDICE B - Tabela dos coeficientes das variáveis do modelo de *Random Forest*

OBESIDADE

Genus	Values
Butyrivibrio	0,210883
Veillonella	0,210381
Anaerotruncus	0,205254
Haemophilus	0,189634
Subdoligranulum	0,183848

DIABETES

Genus	Values
Subdoligranulum	0,221579
Bilophila	0,203845
Parabacteroides	0,202337
Roseburia	0,190899
Peptostreptococcaceae_noname	0,181340