UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Ciência de Dados em Análise de Qualidade de Sequenciamento Sanger

Verônica Maria Rodege Gogola Kolling

Monografia - MBA em Ciência de Dados (CEMEAI)

ICMC USP
SÃO CARLOS

Verônica Maria Rodege Gogola Kolling

Ciência de Dados em Análise de Qualidade de Sequenciamento Sanger

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Afonso Paiva Neto

Versão original

São Carlos

2025

É possível elaborar a ficha catalográfica em LaTeX ou incluir a fornecida pela Biblioteca. Para tanto observe a programação contida nos arquivos USPSC-modelo.tex e fichacatalografica.tex e/ou gere o arquivo fichacatalografica.pdf.

A biblioteca da sua Unidade lhe fornecerá um arquivo PDF com a ficha catalográfica definitiva, que deverá ser salvo como fichacatalografica.pdf no diretório do seu projeto.

Verônica Maria Rodege Gogola Kolling

Data Science in Sanger Sequencing Quality Analysis

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Science

Advisor: Prof. Dr. Afonso Paiva Neto

Original version

São Carlos

2025

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.
No presente modelo consta como
folhadeaprovacao.pdf

Ao meu amado filho, Gabriel

AGRADECIMENTOS

Ao meu esposo, Daniel, pela imensa compreensão e paciência ao longo de muitos meses. Se cheguei até aqui foi porque tive um grande parceiro me incentivando, aguentando as pontas e se desdobrando para que eu tivesse tempo para me dedicar à especialização e a este trabalho.

Aos meus pais, por estimular a busca pelo conhecimento, nutrindo minha curiosidade desde à infância.

Aos meus amigos e familiares mais próximos, obrigada por toda torcida e encorajamento.

Ao meu orientador, professor Afonso Paiva Neto, meu sincero muito obrigada por compartilhar conhecimento e guiar meus passos.

"Na vida, não existe nada a temer, mas a entender."
Marie Curie

RESUMO

KOLLING, V. M. R. G. **Ciência de Dados em Análise de Qualidade de Sequenciamento Sanger**. 2025. 89 p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Sanger é uma técnica de sequenciamento de DNA amplamente utilizada em laboratórios de pesquisa, entretanto resultados de má qualidade impactam diretamente projetos de Biotecnologia. Este trabalho visa modelar os dados operacionais de sequenciadores Sanger, identificando fatores que influenciam na qualidade do sequenciamento. Os dados, extraídos de um \textit{data warehouse} privado, foram divididos em três subconjuntos com base no tipo de amostra sequenciada. 15 algoritmos de classificação foram comparados, abrangendo métodos baseados em árvores de decisão, modelos lineares, abordagens baseadas em instâncias e modelos inspirados em estatística. O desempenho foi avaliado com base no tempo de treinamento e em métricas derivadas matriz de confusão para determinar a capacidade discriminativa e a concordância dos modelos. Dentre estes, os que apresentaram melhor resultado foram XGBoost e LightGBM, ambos baseados em árvores de decisão, com aplicação da técnica de Gradient Boosting para otimização do aprendizado. A metodologia SHAP indicou que os atributos relacionados à quantidade de DNA na amostra, temperatura de anelamento do primer, vida útil dos consumíveis e número de injeções do cartucho de sequenciamento são os mais importantes nas modelagens, o que permitiu sugerir melhorias no protocolo de sequenciamento. É importante destacar que, até a presente momento, não temos conhecimento de referências na literatura de trabalhos com este enfoque.

Palavras-chave: Sequenciamento de DNA. Algoritmos de Classificação. Valores de Shapley. Biotecnologia

ABSTRACT

KOLLING, V. M. R. G. **Data Science in Sanger Sequencing Quality Analysis**. 2025. 89 p. Monograph (MBA in Data Sciences) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Sanger is a widely used DNA sequencing technique in research laboratories; however, poor quality results directly impact Biotechnology projects. This work aims to model the operational data of Sanger sequencers, identifying factors that influence sequencing quality. The data, extracted from a private data warehouse, were divided into three subsets based on the type of sample sequenced. Fifteen classification algorithms were compared, including decision tree-based methods, linear models, instance-based approaches, and statistically inspired models. Performance was evaluated based on training time and metrics derived from the confusion matrix to determine the discriminative capacity and agreement of the models. Among these, the best-performing algorithms were XGBoost and LightGBM, both decision tree-based algorithms that use the Gradient Boosting technique for optimization. The SHAP methodology indicated that attributes related to the amount of DNA in the sample, primer annealing temperature, consumable lifespan, and the number of injections of the sequencing cartridge are the most important in the models, which allowed for suggesting improvements in the sequencing protocol. It is important to highlight that, to date, we have not found any literature references that focus on this specific topic.

Keywords: DNA Sequencing. Classification Algorithms. Shapley Values. Biotechnology

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A	Adenina
AUC	Área sob a Curva
C	Citosina
CRL	do inglês, Continuous Reading Length
DBTL	do inglês, Design - Build - Test - Learn
DNA	Ácido desoxirribonucleico
dNTP	Desoxirribonucleotídeo
ddNTP	Dideoxirribonucleotídeo
G	Guanina
kNN	do inglês, k-nearest neighbours
LightGBM	do inglês, Light Gradient-Boosting Machine
MCC	Coeficiente de correlação de Mattews
NGS	Sequenciamento de Nova Geração
PCR	Reação em cadeia da polimerase
QS	do inglês, Quality Score
QV	do inglês, Quality Value
QV20+	Quality Value acima de 20
ROC	Característica de Operação do Receptor
SHAP	do inglês, SHapley Additive exPlanations

SVM	Máquina de Vetores de Suporte
T	Timina
Tm	Temperatura de anelamento
TT	Tempo total de treinamento
XGBoost	do inglês, Extreme Gradient Boosting

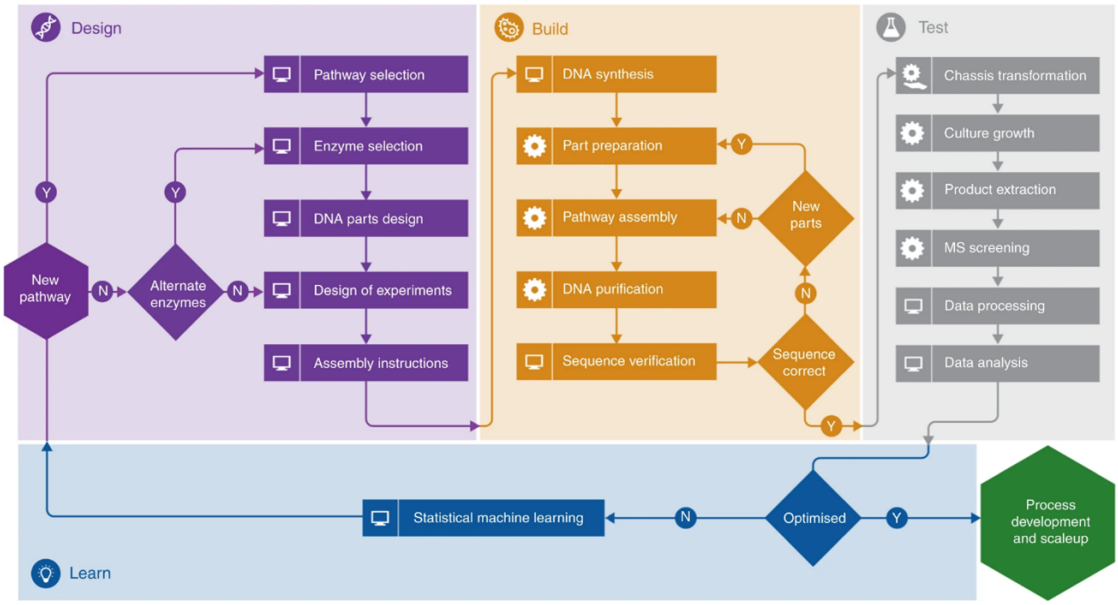
SUMÁRIO

1 INTRODUÇÃO

A engenharia de genética possui um papel fundamental na Biotecnologia Industrial, permitindo o desenvolvimento de organismos geneticamente modificados para aplicações diversas, como a produção de bioquímicos, fármacos, químicos, combustíveis, enzimas e terapias biológicas (Nielsen; Keasling, 2016). Para que essas aplicações sejam viáveis, é essencial garantir que as modificações genéticas sejam corretamente implementadas no organismo de interesse, o que torna a validação genética uma etapa crítica dentro dos fluxos de trabalho em laboratórios de biotecnologia.

O fluxo de trabalho iterativo DBTL (do inglês, Design - Build - Test - Learn), representado na Figura 1 é comumente utilizado em projetos de Biotecnologia Industrial.

Figura 1 – Workflow de projetos de biotecnologia



Fonte: Carbonell *et al.* (2018).

Neste ciclo, uma nova aplicação segue pelas etapas de: (i) desenho das modificações genéticas; (ii) construção e obtenção do organismo geneticamente modificado; (iii) teste do organismo recém-obtido para a aplicação de interesse; e (iv) aprendizado a partir dos testes para desenho de novas modificações genéticas, retornando à etapa (i) até que se obtenha um organismo adequado para industrialização (Carbonell *et al.*, 2018).

Na etapa (ii) ocorre a validação genética, ou seja, a confirmação de que as células do organismo engenheirado possuem a sequência de DNA idêntica à desenhada na etapa (i). Sem essa confirmação, não é possível prosseguir para a etapa seguinte de teste (etapa iii). Para a confirmação, uma amostra de DNA do organismo é sequenciada com um equipamento específico chamado sequenciador. A sequência de DNA obtida é comparada com a sequência teórica. Por isso, é primordial que a qualidade do sequenciamento seja satisfatória para que a comparação ocorra com exatidão e o projeto de Biotecnologia possa ser desenvolvido.

O *sequenciamento Sanger* (Sanger; Nicklen; Coulson, 1977), também conhecido como método de terminação de cadeia, é uma das possíveis técnicas de sequenciamento e foi descrito pela primeira vez no final da década de 70. Amplamente utilizado, atualmente o método automatizado do sequenciamento Sanger consta de 4 etapas principais: (a) obtenção da amostra de DNA que se deseja sequenciar em pureza adequada; (b) execução da reação de PCR de terminação de cadeia, na qual serão obtidas bilhões de cópias da sequência de DNA da amostra, que são fragmentos terminados em tamanhos aleatórios por um nucleotídeo fluorescente; (c) organização dos fragmentos do menor para o maior tamanho, detecção da fluorescência terminal dos fragmentos pelo sensor do sequenciador e conversão das fluorescências detectadas para sequência de nucleotídeos; e (d) comparação da sequência de nucleotídeos obtida com a sequência original usando um software específico de alinhamento (Heather; Chain, 2016). No sequenciamento Sanger, a qualidade do resultado é multifatorial e depende da característica da amostra, da sequência de DNA em si, da manipulação do operador e da operação do equipamento (Life Technologies, 2016).

Em um dado laboratório de Biotecnologia (empresa "X"[1]), cerca de 30% das

[1] O nome da empresa será omitido por questões de sigilo

amostras sequenciadas em sequenciador do tipo Sanger estão classificadas como de baixa qualidade. Sequenciamentos de baixa qualidade impactam no desenvolvimento de projetos de Biotecnologia em termos prazo, competitividade e custos. De acordo com Crossley *et al.* (2020), taxas de falhas de sequenciamento acima de 10% devem ser investigadas para se determinar qual a principal interferência na qualidade do sequenciamento.

Neste contexto, propõe-se analisar e modelar os dados de operação dos sequenciadores do tipo Sanger de forma a identificar fatores que devem ser considerados pelos operadores durante a execução do sequenciamento, de forma a diminuir a taxa de amostras com baixa qualidade. A abordagem proposta se enquadrada dentro dos pilares de efetividade industrial, no qual são almejadas soluções de baixo investimento e eficientes para redução de custos e aumento de competitividade.

1.1 Objetivos

O objetivo geral deste trabalho é modelar os dados de operação de sequenciadores do tipo Sanger de um laboratório de Biotecnologia e identificar quais mudanças no protocolo de obtenção de amostra de DNA e/ou de operação do equipamento devem ser feitas para diminuir a taxa de resultados de sequenciamento com baixa qualidade.

Os seguintes objetivos específicos são propostos:

1. Estudar e entender principais aspectos da operação do equipamento e como os dados “crus” do sensor do equipamento são convertidos para índices de qualidade;

2. Diminuir o número de atributos que serão incluídos na modelagem, através de estudo na literatura em conjunto com aprendizado obtido no objetivo (1);

3. Estudar modelos de classificação frequentemente aplicados em problemas de biologia sintética;

4. Comparar a performance dos modelos de classificação;

5. Identificar, entre os modelos com boa predição, quais atributos têm maior impacto no índice de qualidade de sequenciamento;

6. Sugerir mudanças no protocolo de obtenção de amostra de DNA e/ou de operação do equipamento e avaliar os resultados potenciais da implementação das mudanças propostas.

1.2 Estrutura do Trabalho

Este trabalho está divido em 5 capítulos principais. Em sequência à introdução, no Capítulo 2 será apresentada uma fundamentação teórica sobre o DNA e o Sequenciamento Sanger, conceitos essenciais pra o entendimento do trabalho, e também uma revisão da literatura sobre ciência de dados aplicada ao controle de qualidade de dados de sequenciamento de DNA. Ao final desse capítulo serão apresentados algoritmos comumente aplicadas em tarefas de classificação. No Capítulo 3 serão detalhadas as etapas para a modelagem dos dados, incluindo uma analise descritiva e exploratória dos dados. Logo depois, no Capítulo 4, são descritos os resultados da modelagem e suas análises. Por fim, no Capítulo 5 apresenta-se um resumo do trabalho assim como algumas observações de melhorias e trabalhos futuros.

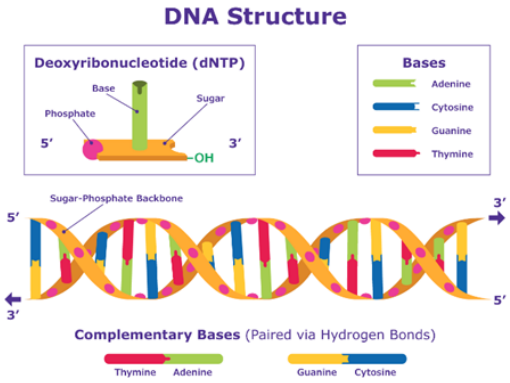
2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO BIBLIOGRÁFICA

Na primeira seção deste capítulo serão apresentados conceitos fundamentais para entendimento do sequenciamento Sanger e seus dados. A segunda seção compreende uma revisão bibliográfica das principais publicações sobre ciência de dados aplicada a controle de qualidade de dados das principais técnicas de sequenciamento de DNA. Na terceira seção serão discutidos os principais algoritmos de classificação relevantes para este trabalho.

2.1 O DNA, sua estrutura e a Reação em Cadeia da Polimerase

O DNA é uma molécula composta de duas fitas. Cada fita é composta por uma cadeia de moléculas chamadas *desoxirribonucleotídeos* (dNTPs). Cada dNTP contém um grupo fosfato, um grupo açúcar e uma das quatro bases nitrogenadas: *adenina* (A), *timina* (T), *guanina* (G) ou *citosina* (C). Para formar a dupla fita de DNA, dNTPs entre duas fitas são unidos por ligações de hidrogênio entre bases complementares. Como resultado, as fitas se enrolam uma em torno da outra para formar estrutura de dupla hélice, conforme representado na Figura 2 (**??**).

Figura 2 – Estrutura do DNA

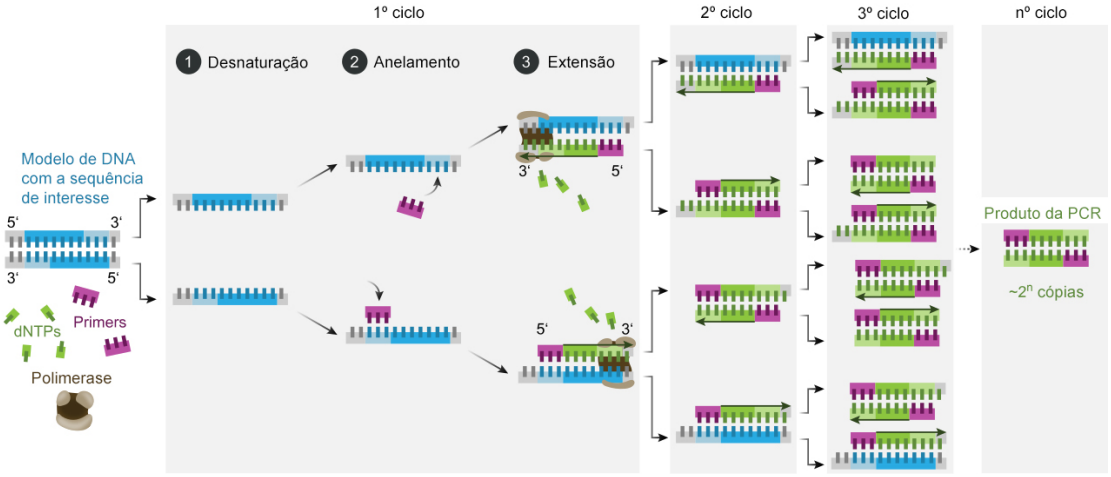


Fonte:**??**)

A *reação em cadeia da polimerase* (PCR) é uma técnica rotineira de labora-

tório usada para fazer muitas cópias (milhões ou bilhões) de uma região específica do DNA, ilustrada na Figura 3.

Figura 3 – Reação em cadeia da polimerase – PCR



Fonte: ??)

Assim como a replicação de DNA em um organismo, a PCR requer uma enzima chamada DNA polimerase que faz novas fitas de DNA usando as existentes como moldes. Na PCR, a DNA polimerase adiciona dNTPs livres em solução a uma fita molde de DNA em crescimento, catalisando a formação de uma nova fita de DNA. A DNA polimerase consegue fabricar DNA somente quando lhe é dada uma sequência curta de nucleotídeos que fornece um ponto de partida para a síntese de DNA, chamado de primer. Os primers de PCR são pedaços curtos de DNA de fita simples, geralmente por volta de 20 nucleotídeos de comprimento. Quando os primers são ligados ao molde, eles podem ser estendidos pela polimerase, e a região que está entre eles será copiada (??)

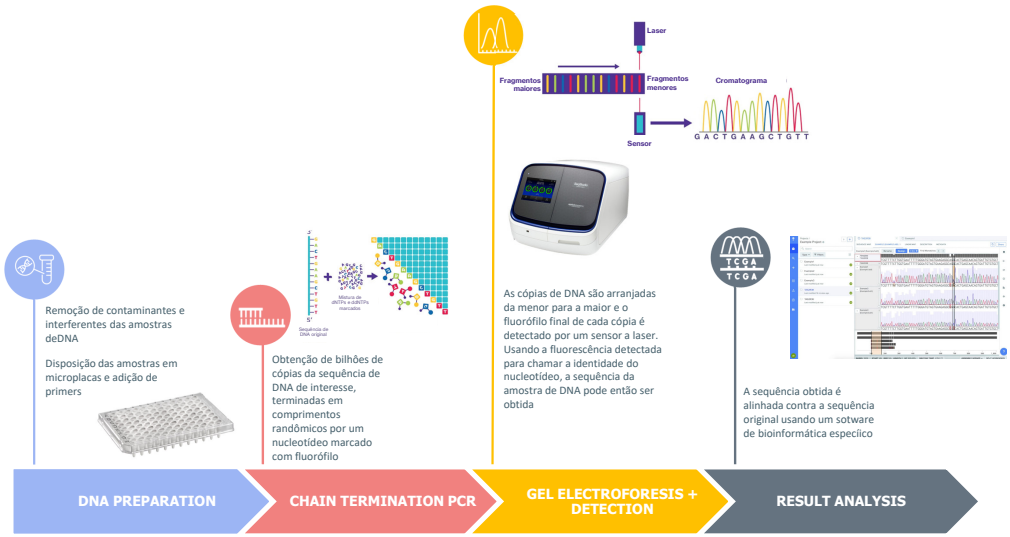
2.2 Sequenciamento Sanger

O sequenciamento Sanger (Sanger; Nicklen; Coulson, 1977), também conhecido como método de terminação de cadeia, é uma das possíveis técnicas de sequenciamento de DNA e é amplamente utilizado. Apesar de novas e avançadas técnicas de sequenciamento de DNA serem utilizadas atualmente, o sequenciamento

Sanger continua sendo uma técnica relevante e utilizada mundialmente por sua acurácia diferenciada, custo-benefício para aplicações específicas, requer pouca quantidade de DNA e tem uma metodologia bem estabelecida (**??**).

Atualmente o método automatizado consta de 4 etapas principais, conforme ilustrado na Figura 4 e descrito em maiores detalhes nos tópicos a seguir:

Figura 4 – Etapas envolvidas no Sequenciamento Sanger



Fonte: Elaborado pelo autor.

2.2.1 Etapa 1: obtenção da amostra de DNA que se deseja sequenciar em pureza adequada

Esta etapa é crítica e determinante na qualidade final do sequenciamento. Para o sequenciamento, é essencial que a amostra de DNA esteja livre de contaminantes e impurezas e que o DNA esteja íntegro. Existem diversos protocolos e kits comerciais que isolam o DNA de matrizes complexas em condições adequadas para o sequenciamento. Entretanto, oscilações de rendimento e de pureza da amostra podem ocorrer de acordo com a natureza da matriz da qual o DNA será isolado, do tipo do DNA que será isolado e do operador. Em um trabalho de comparação de DNA isolado utilizando diferentes kits no resultado de sequenciamento, foi

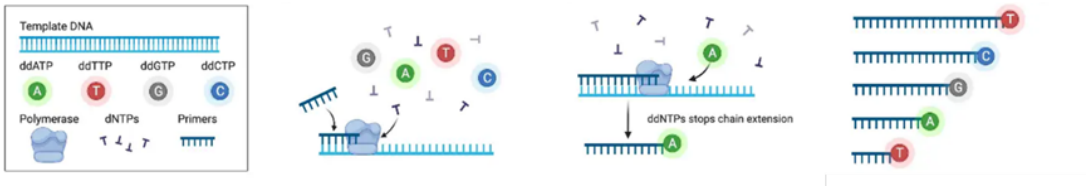
verificado que a quantidade de proteína residual no DNA purificado pode interferir na qualidade do sequenciamento (**??**).

Após a purificação do DNA, o DNA é quantificado e uma fração do DNA é transferido para microplacas em quantidades apropriadas para o sequenciamento. As quantidades de DNA são sugeridas pelo fabricante do sequenciador. Nesta etapa é adicionado um primer ao DNA para que a PCR diferenciada possa acontecer. Se é desejado que uma amostra seja sequenciada com mais de um primer (n primers), a amostra deverá ser transferida para n poços da microplaca e em cada poço será adicionado um primer em específico. A escolha do primer é essencial para o sucesso do sequenciamento e o primer deve atender uma série de especificações, dentre elas temperatura de anelamento (Tm) (Crossley *et al.*, 2020) .

2.2.2 Etapa 2: execução da reação de PCR de terminação de cadeia

A sequência de DNA de interesse é usada como modelo para um tipo especial de PCR chamado PCR de terminação de cadeia (Figura 5). A PCR com terminação em cadeia funciona como a PCR padrão, mas com uma grande diferença: na PCR de terminação de cadeia, o usuário mistura uma baixa proporção de ddNTPs com os dNTPs normais na reação de PCR. Os ddNTPs são nucleotídeos modificados chamados de *dideoxirribonucleotídeos*. Quando a DNA polimerase incorpora um ddNTP aleatoriamente, a extensão da fita em crescimento é interrompida. O resultado da PCR de terminação de cadeia é de milhões a bilhões de cópias da sequência de DNA de interesse, terminadas em um comprimento aleatório por ddNTPs. A razão de ddNTP e dNTP na reação de PCR de terminação de cadeia é otimizada para obter uma população balanceada de fragmentos curtos e longos (**??**). No Sanger automatizado, cada um dos ddNTPs (ddA, ddG, ddC, ddT) é marcado com uma molécula fluorescente específica, permitindo sua detecção na eletroforese capilar (etapa 3).

Figura 5 – PCR de terminação de cadeia



Fonte: adaptado de **??**).

Resíduos da reação PCR de terminação de cadeia afetam a qualidade do sequenciamento. Portanto, após a PCR de terminação de cadeia é executada uma etapa de purificação para remoção de primers, dNTPs e ddNTPs não incorporados, redução de concentração de sais, tampão e outros contaminantes (Life Technologies, 2016).

2.2.3 Etapa 3: Eletroforese Capilar e Detecção

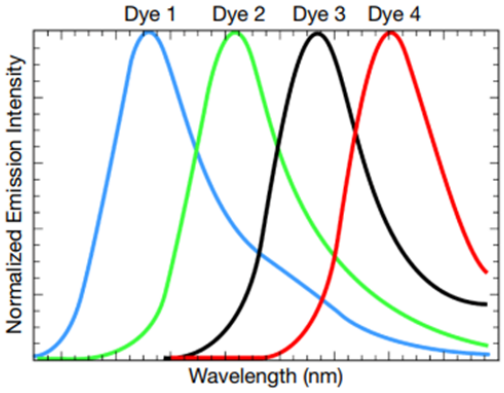
Na terceira etapa, os fragmentos obtidos na etapa 2 são separados por meio da eletroforese capilar em gel, que é executada no equipamento de sequenciamento. Na eletroforese capilar em gel, o DNA é injetado em uma extremidade de uma matriz de gel e uma corrente elétrica é aplicada. O DNA é naturalmente carregado negativamente, portanto, os fragmentos de DNA serão puxados em direção ao eletrodo positivo, no lado oposto do gel. A velocidade de movimentação dos fragmentos é determinada pelo seu tamanho: quanto menor for um fragmento, menos atrito ele sofrerá ao se mover pelo gel e mais rápido ele se moverá. Como resultado, os fragmentos serão organizados do menor para o maior com uma resolução de um dNTP; ou seja, a eletroforese é capaz de ordenar os milhões de fragmentos obtidos na reação de PCR de terminação de cadeia, retardando a migração de fragmentos que tenham um dNTP a mais que o fragmento anterior (Life Technologies, 2016).

Como cada fragmento tem um ddNTP terminal marcado, quando este fragmento passar pela janela de detecção próxima ao eletrodo positivo, emitirá um sinal. O gráfico obtido ao representar a intensidade de sinal de cada um dos sinais fluorescentes detectados é chamado de cromatograma. Portanto, ao interpretar o

cromatograma, considerando o padrão de migração de fragmento do menor para o maior e que cada um dos quatro ddNTPs é marcado com uma etiqueta fluorescente diferente, o sinal detectado pelo sequenciador pode ser diretamente associado à identidade do ddNTP terminal, processo denominado *basecalling* (**??**).

Durante o *basecalling*, algoritmos são empregados para corrigir eventuais distorções de mobilidade entre os ddTNPs e para determinar a exatidão da chamada de base. Cada ddNTP emite uma fluorescência máxima em um determinado comprimento de onda, porém há uma sobreposição de emissão (Figura 6). Ou seja, um sinal gerado primariamente em um canal de cor irá ter um sinal menor um canal de cor adjacente (**??**).

Figura 6 – Espectro de emissão de 4 fluorófilos comumente utilizados em sequenciamento Sanger



Fonte: **??**).

Portanto, no *basecalling* o sinal "principal" e o sinal de fundo (*background*) são analisados e um valor de qualidade da predição, do inglês *quality value* (QV), para cada base é calculado da seguinte forma:

$$QV = -10 \times \log(\text{probabilidade de erro})\,. \tag{2.1}$$

O valor QV indica a probabilidade de erro na chamada de base (Tabela 1). Por exemplo, um QV de 20 prediz uma taxa de erro de 1% enquanto que um QV de

40 indica que as chances da base ser chamada incorretamente é de apenas 1 em 10.000 (**??**).

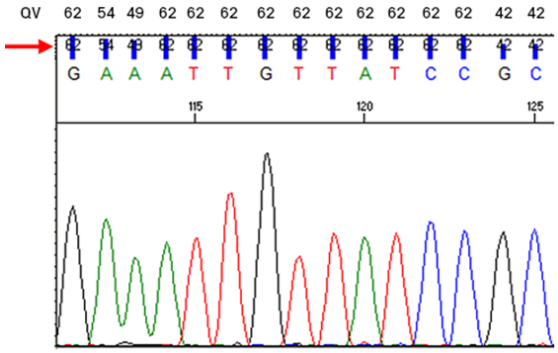
Tabela 1 – Acurácia da chamada de base de acordo com o QV calculado

QV	**Acurácia da chamada de base (%)**
10	90
20	99
30	99,9
40	99,99
50	99,999

Fonte: Fonte: **??**).

A saída processada do *basecalling* é chamado de *trace file* (Figura 7), composto de quatro séries temporais de sinais. Em outras palavras, o *trace file* contém os dados brutos do sequenciamento e também os valores de QV para as bases reconhecidas.

Figura 7 – Exemplo de trace file: cromatograma de Sequenciamento Sanger



Legenda: no *trace file* os valores de QV e a identificação da base são ilustrados juntamente com os sinais detectados dos fluorófilos ao longo do tempo. Fonte: **??**).

A técnica atual de sequenciamento Sanger geralmente suporta a geração de cromatogramas com 800-1000 bases. Tipicamente, as primeiras 15 a 40 bases não são bem resolvidas e não geram bases com boa qualidade. Após as 500 bases são esperados picos menos com menor definição e o *basecalling* é menos confiável.

A maior parte dos protocolos de sequenciamento é otimizada para prover uma resolução ótima entre 100 e 500 bases (*clear range*). Nessa faixa os picos estão bem espaçados, com boa simetria e o *basecalling* é confiável (**??**); Crossley *et al.* (2020)).

Além do QV para cada base no *basecalling*, os equipamentos de sequenciamento tipo Sanger em geral fornecem algumas métricas adicionais de qualidade geral dos *trace files* que podem dar uma noção da qualidade do sequenciamento.

2.2.3.1 Métricas de Qualidade

Na Tabela 2, são descritas as 7 métricas fornecidas pelo equipamento Sanger utilizado no laboratório de Biotecnologia da empresa "X"ao final do sequenciamento de uma amostra de DNA.

Com as métricas, o equipamento atribui uma categoria para o resultado de sequenciamento:

- Verde: todas as métricas estão acima do limite superior;
- Amarelo: ao menos uma das métricas está entre limite superior e inferior; e
- Vermelho: ao menos uma das métricas está abaixo do limite inferior.

Tabela 2 – Métricas de qualidade de amostras de DNA sequenciadas

Métrica	**Descrição**	**Limite inferior**	**Limite superior**
QV20+	Número de bases total com QV >=20	QV20+ < 100 são consideradas amostras de baixa qualidade	QV20+ > 300 são consideradas amostras de alta qualidade
CRL (Continuous Reading Length)	Contagem máxima de número de bases em sequência com QV >=20 considerando a média móvel de uma janela de 20 bases	CRL > 180 são consideradas amostras de alta qualidade	CRL > 510 são consideradas amostras de alta qualidade
QS (Trace Score ou Quality Score)	QV médio na faixa com boa leitura (clear range)	QS < 20 são propensos a baixo sinal e/ou alto ruído.	QS >30 tem boa qualidade
Median PuP	Mediana da razão entre o sinal do pico principal e o sinal do pico secundário dentro da clear range	<10, alto ruido	> 15 pouco ruído
Signal strength	Média da fluorescência relativa dos quatro fluorófilos no cromatograma	<200 sinal baixo	>600 bom
Signal to noise	Média de fluorescência relativa de sinal-ruído para os quatro fluorófilos no cromatograma	<60 muito ruído	>180 baixo ruído
Offscale	Pelo menos um ponto de dados na clear range saturou o sensor.	Se 1 ponto saturou o sensor é sinal de corrida com alto ruído	não se aplica

Fonte: **??**), **??**).

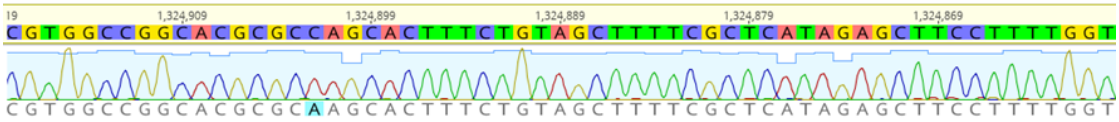
Embora existam diretrizes para verificar se um resultado de sequenciamento está dentro ou fora dos limites considerados “bons”, não existe um score único para definir a qualidade geral do sequenciamento. **??**) demonstraram que uma base de dados de sequenciamento é muito heterogênea para se utilizar apenas uma métrica para predizer a qualidade de um resultado de sequenciamento. Os autores

propõem uma combinação das métricas QV e *signal strength* juntamente com a inspeção visual para determinar se um resultado de sequenciamento é confiável para ser utilizado em aplicações clínicas. **??**), por outro lado, propõem novas métricas similares a CRL para determinar a qualidade de um sequenciamento. **??**), defendem que criar um score único combinando diferentes métricas não é uma prática que deve ser adotada, uma vez que cada base chamada contém um elemento de erro cuja distribuição é incerta.

2.2.4 Etapa 4: Análise dos Resultados

A sequência obtida no sequenciador através dos *trace files* é comparada com a sequência teórica desenhada utilizando softwares específicos que, em geral, não são relacionados ao sequenciador (Figura 8). Nessa etapa de alinhamento, a sequência do *tracefile* é comparada nucleotídeo por nucleotídeo com a sequência teórica, portanto é primordial que a qualidade do sequenciamento seja satisfatória para fazer uma comparação segura. Apesar dos softwares fazerem um bom trabalho de comparação entre as bases, o alinhamento não é isento de erro e por isso se faz necessária uma validação cruzada manual (**??**). Nesta etapa, o pesquisador decidirá se a qualidade do sequenciamento está satisfatória, implicando na necessidade ou não de repetir o sequenciamento.

Figura 8 – Exemplo de alimento de um trace file com sequência teórica



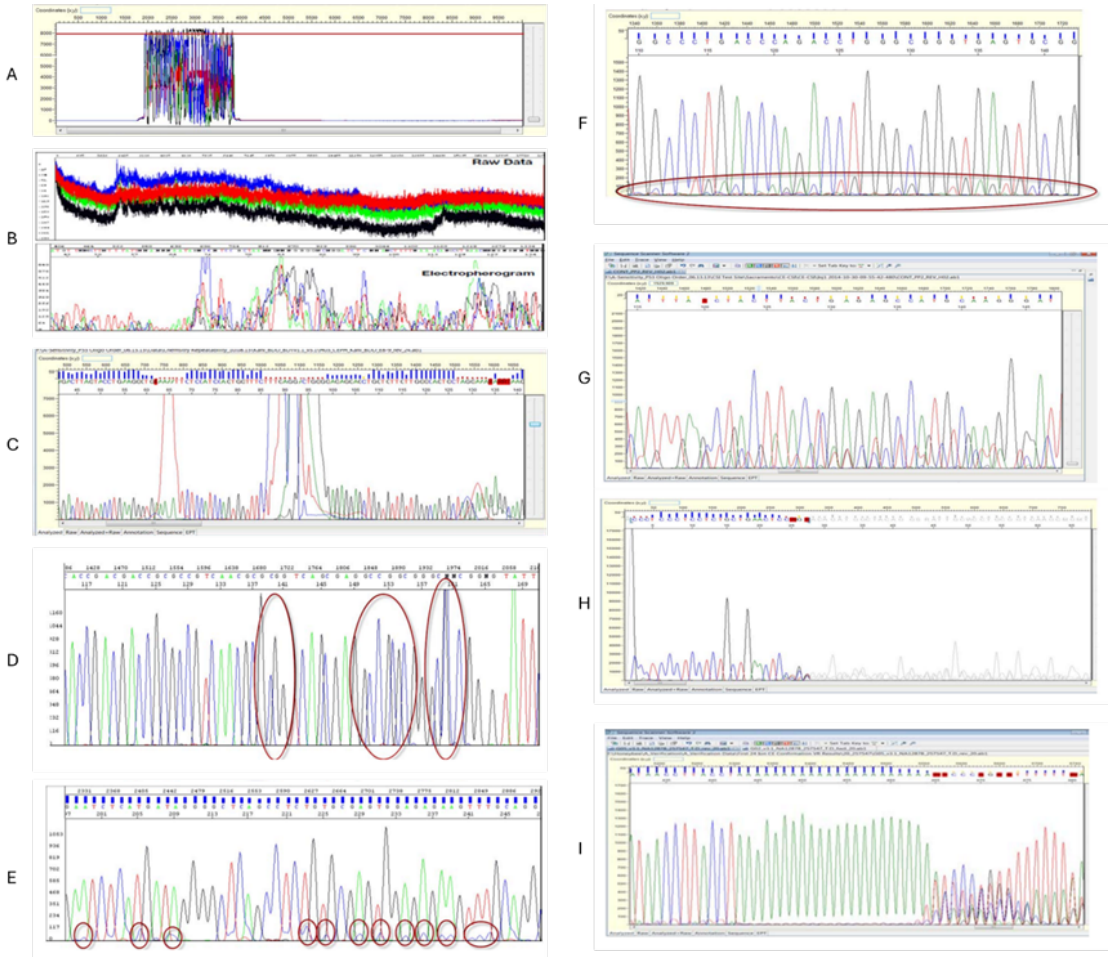
Fonte: elaborado pelo autor.

2.2.4.1 Fontes de Resultados de Sequenciamento com má qualidade

Há muitas razões pelas quais um Sequenciamento do tipo Sanger pode apresentar má qualidade (Life Technologies (2016); **??**)). Os principais são ilustrados na Figura 9 e descritos a seguir:

- Fita molde tem regiões ricas com bases G e C, que causam compressão dos dados;
- Formação de sequências palindrômicas ou formação de sequências com alto grau de estrutura secundária;
- Presença de homopolímeros;
- Quantidade de DNA utilizado e proporção de primer utilizada na reação de PCR de terminação de cadeia. Muito DNA pode apresentar picos muito intensos no início do cromatograma que diminuem muito rapidamente. Pouco DNA pode apresentar sinal com baixa intensidade e muito ruído;
- Qualidade da DNA. DNA degradado irá gerar resultados de sequenciamento com menor qualidade;
- Presença de contaminantes remanescentes da extração de DNA e depois na purificação pós reação de PCR de terminação de cadeia;
- Contaminantes no primer;
- Contaminação cruzada de amostras;
- Degradação dos fluorófilos;
- Amplificações não-específicas na reação de PCR de terminação de cadeia;
- Condição geral dos componentes consumíveis do sequenciador, como a matriz de gel e tampão utilizado na eletroforese; e
- Falha na operação do equipamento (falha na injeção, oscilação em corrente elétrica).

Figura 9 – Exemplos de sequenciamento Sanger com má qualidade



Legenda: (A) excesso de DNA na amostra, (B) Baixa concentração de DNA na amostra; DNA não íntegro; componentes consumíveis do sequenciador em condições inadequadas para sequenciamento, (C) Presença de contaminantes, (D) Fita molde tem regiões ricas com bases G e C, (E) Degradação dos fluorófilos, (F) Contaminantes no primer, (G) Contaminação cruzada de amostras, (H) amplificações não-específicas, e (I) Presença de homopolímeros. Fonte: adaptado de Life Technologies (2016).

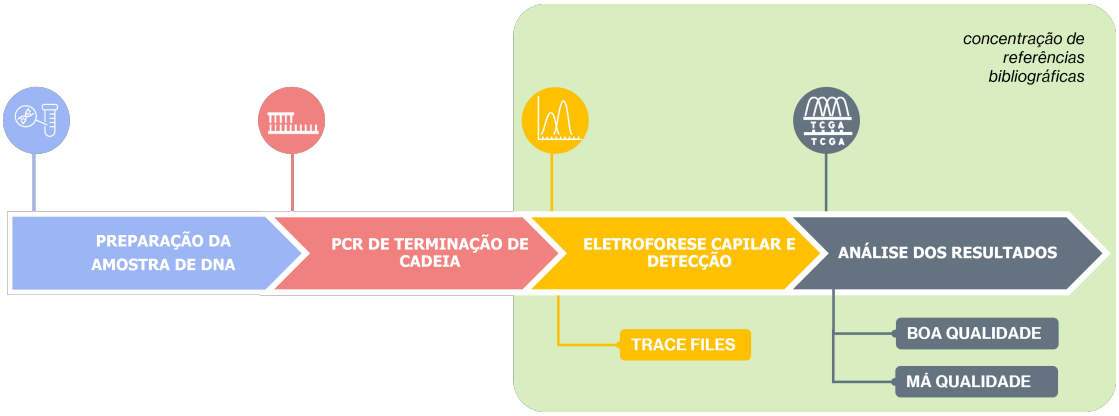
2.3 Ciência de Dados aplicada a controle de qualidadade de dados de Sequenciamento de DNA

Com o advento e disseminação das técnicas de sequenciamento de DNA, o volume de dados referente a sequências de DNA cresceu exponencialmente nas

últimas décadas. O aprendizado de máquina é uma técnica poderosa utilizada amplamente por bioinformatas para extrair informações de dados biológicos e que tem sido decisiva no avanço de inúmeras pesquisas. Entre as principais frentes no aprendizado de máquina utilizando dados de DNA é possível citar: alinhamento de sequências de DNA, classificação de sequências de DNA, clustering de sequências de DNA e mineração de padrões de sequência de DNA. Essas frentes podem exploradas em mais detalhes nas revisões de **??**) e **??**).

Além das frentes principais acima citadas, no estado da arte da ciência de dados de sequenciamento de DNA, se fazem presentes referências para correlação entre os dados brutos de qualidade de sequenciamento e classificadores gerais de qualidade (Figura 10). Por outro lado, há uma deficiência de estudos de ciência de dados focados no entendimento dos fatores que influenciam o sequenciamento, ou seja: há uma lacuna na literatura de estudos que correlacionem sistematicamente os fatores e parâmetros desde o momento de preparo das amostras de DNA até a análise dos resultados. Embora estudos isolados de centros de sequenciamento pontuem sobre potenciais fontes de sequenciamento de má qualidade (**??**) e de desempenho geral de *facility* (**??**), não foram encontrados estudos preditivos correlacionando informações de pré-sequenciamento com a qualidade final do sequenciamento, objetivo principal deste trabalho.

Figura 10 – Estado da arte em ciência de dados de sequenciamento



Fonte: Elaborado pelo autor.

Os principais trabalhos encontrados na área de ciência de dados de sequenciamento de DNA com enfoque em controle de qualidade estão sumarizados na Tabela 3. Todos os trabalhos citados têm em comum a motivação de retirar do pesquisador a avaliação se um sequenciamento foi bom ou ruim e agilizar a seleção de dados de sequenciamento confiáveis para as posteriores análises. Em todos os trabalhos, um modelo foi treinado para predizer a qualidade de sequenciamento com base em dados extraídos dos *trace files*.

Tabela 3 – Publicações referência em ciência de dados de sequenciamento com enfoque em controle de qualidade

Referência	Dados	Atributos	Modelos avaliados	Critérios de avaliação	Melhor modelo
Öz e Kaya, 2013	48 trace files de sequenciamento tipo Sanger	8 atributos obtidas por estatística descritiva a partir dos sinais dos trace files	Máquinas de vetores de suporte	matriz de confusão	Máquinas de vetores de suporte. Acerto de 23 de 24 predições
Öz et al., 2016	2301 trace files de sequenciamento tipo Sanger	15 atributos obtidas por estatística descritiva a partir dos sinais dos trace files	• k-vizinhos mais próximos • Naive Bayes • Máquinas de vetores de suporte com kernels de função de base radial • Máquinas de vetores de suporte com kernels de função de polinomial • Florestas aleatórias	• matriz de confusão, • coeficiente kappa, • erro médio absoluto, • taxa de falso positivos, • acurácia e precisão, • recall ou sensitividade, • F-score, • coeficiente de correlação de Matthew, • tempo para construir o modelo	Florestas aleatórias acurácia 94,68% 4 atributos
Kurt et al., 2018	2301 trace files de sequenciamento tipo Sanger	7 atributos obtidas por estatística descritiva a partir dos sinais dos trace files	• Reptree • J48 (C4.5) • Florestas aleatórias • Regressão Logística	• Acurácia, • erro quadrático médio, • f-score	Florestas aleatórias acurácia 95,26% 5 atributos
Albrecht et al., 2021	2642 dados de sequenciamento tipo NGS	11 atributos relacionados aos dados brutos e 27 atributos resultantes do alinhamento do sequenciamento contra a sequência alvo	• Florestas aleatórias • Redes neurais • Gradient boosting • k-vizinhos mais próximos • XG-boost • Máquinas de vetores de suporte • Naive Bayes • Regressão Logística • Árvores Extremamente Aleatórias • Árvore de decisão	Curva ROC	Florestas aleatórias auROC = 0,925
Sprang et al., 2021	2098 dados de sequenciamento tipo NGS	12 atributos relacionados aos dados brutos e 35 atributos resultantes do alinhamento do sequenciamento contra a sequência alvo	Árvores de decisão	• acurácia, • precisão, • recall ou sensitividade, • F-score	Árvores de decisão, acurácia média de 0,987

Em dados de sequenciamento Sanger, o grupo de Öz e colaboradores publicaram 3 artigos relacionando a predição de qualidade de sequenciamento com dados obtidos a partir dos *trace files*. Em 2013, o grupo publicou um artigo no qual se buscava responder se o modelo baseado em Máquinas de vetores de suporte teria potencial para ser utilizado como classificador de dados de sequenciamento. Neste estudo, foram utilizados apenas 48 dados de sequenciamento e o modelo foi avaliado utilizando somente a matriz de confusão e utilizando 8 atributos calculados por estatística descritiva a partir dos sinais dos *trace files*. Os 8 atributos foram: média, desvio padrão e mediana de todos os valores no cromatograma; média, desvio padrão e mediana dos valores de pico em cada um dos quatro canais no cromatograma; número de picos no cromatograma; e média do pico do sinal formado pela soma dos valores de cada canal no cromatograma. Foi verificado boa performance do modelo, que falhou em 1 de 24 predições do conjunto teste (**??**).

Já **??**) ampliaram a base de dados, passando a utilizar um banco de dados público de sequenciamento. O grupo manualmente curou 2301 *trace files*, classificando-os como: 1626 sequências com alta qualidade e 631 sequências de baixa qualidade. O grupo deixou de fora do modelo 14 sequências classificadas como "dados sem significado" e 30 sequências como "indecisas" (apresentavam trechos com alta qualidade e trechos com baixa qualidade). A partir dos sinais dos trace files das 2257 sequências, os pesquisadores extraíram dados de 15 atributos, agrupados em 3 famílias:

- família x: dados agrupados dos 4 sinais de *basecalling*;
- família y: dados agrupados da probabilidade de sinal de cada canal; e
- família z: soma dos valores de probabilidade dos 4 sinais

Para cada família foram obtidos os valores de média, mediana, desvio padrão, assimetria e curtose, totalizando 15 atributos. 5 modelos de aprendizado de máquina foram aplicados: k-vizinhos mais próximos, Naive Bayes, Máquinas de vetores de suporte com kernels de função de base radial, Máquinas de vetores de suporte com kernels de função de polinomial e Florestas aleatórias. Foi utilizada validação cruzada com 10 grupos e para avaliar a performance dos modelos foram utilizadas as

seguintes métricas: matriz de confusão, coeficiente kappa, erro médio absoluto, taxa de falso positivos, acurácia, precisão, revocação ou sensitividade, f-score, coeficiente de correlação de Matthew e tempo para construir o modelo. Todos os classificadores tiveram uma ótima performance considerando as 15 atributos (acurácia $> 93\%$). Os 15 atributos foram ranqueados considerando seu poder de discriminação e os top 3, 4 e 5 atributos da lista foram selecionadas para comparar os modelos originais com classificadores simplificados. Os pesquisadores determinaram que utilizando 4 atributos é possível ter uma boa performance com menor recurso computacional. Os 4 atributos são: média e desvio padrão dos 4 sinais agrupados e curtose e desvio padrão da soma dos valores de probabilidade dos 4 sinais. O melhor modelo foi Florestas aleatórias com acurácia de 94,68%, kappa 0,8679 e erro absoluto médio de 0,0722.

Em 2018 (**??**), o grupo continuou trabalhando com o mesmo conjunto de dados de 2016 e voltou a utilizar 7 dos 8 atributos estudados em 2013. Para este estudo, os seguintes modelos de aprendizado de máquina foram utilizados: Reptree, J48 (C4.5), Florestas aleatórias e Regressão Logística. O grupo manteve a validação cruzada com 10 grupos e as métricas para avaliação dos modelos foram: acurácia, erro quadrático médio e f-score. Novamente todos os classificadores tiveram uma ótima performance considerando os 7 atributos (acurácia $> 93\%$). Os 7 atributos foram ranqueados considerando seu poder de correlação com a variável resposta e os 4 atributos com valor de correlação maior que 0,5 foram selecionadas para comparar os modelos originais com classificadores simplificados. Os pesquisadores novamente encontraram que utilizando 4 atributos é possível ter uma boa performance com menor recurso computacional. Os 4 atributos foram: mediana de todos os valores no cromatograma, a média e o desvio padrão dos valores de pico em cada um dos quatro canais no cromatograma e a média do pico do sinal formado pela soma dos valores de cada canal no cromatograma. O melhor modelo foi Florestas aleatórias utilizando AdaBoost com acurácia de 94,90%, f-score 0,949 e erro quadrático médio de 0,225. Foi também avaliado o p-valor das atributos no modelo gerado por Regressão Logística e verificado que além dos 4 atributos já identificados, 1 atributo também tinha significância: o desvio padrão de todos os valores no cromatograma. Os modelos continuaram com ótima predição considerando os 5 atributos. O melhor

modelo considerando 5 atributos foi novamente Florestas aleatórias utilizando AdaBoost com acurácia de 95,26%, f-score 0,953 e erro quadrático médio de 0,216. Este último estudo, entretanto, foi retratado pela revista Neural Computing and Applications em 2024. Foi verificado que nos artigos 2016 e 2018 os dados estavam desbalanceados para uma das classes e não foi feita uma validação externa ou com outros dados/subgrupos para verificar a generalidade do modelo e possíveis vieses. Também não foram descritos os parâmetros utilizados dentro de cada modelo aplicado.

Já para dados de NGS, uma técnica alternativa e mais moderna para o sequenciamento Sanger, se destaca o trabalho de **??**). O grupo escolheu 2642 dados de sequenciamentos, igualmente divididos entre alta qualidade e má qualidade. 10 algoritmos diferentes foram avaliados, nos quais os hiperparâmetros foram otimizados utilizando a ferramenta Grid Search do Scikit-Learn. Foi utilizada validação cruzada com 10 grupos e a métrica de performance escolhida para comparação dos modelos foi a curva ROC. Considerando todos os atributos (11 atributos relacionados aos dados brutos e 27 atributos resultantes do alinhamento do sequenciamento contra as sequências alvo), foi alcançado um valor de curva ROC de 0,925 utilizando o algoritmo de Florestas aleatórias. Para verificar a generalidade do modelo, o grupo testou o modelo contra subconjuntos específicos dos dados originais e obteve bons resultados (ROC $>$ 0,9), o que significa que o modelo não foi enviesado. Para expandir a aplicação do modelo, o grupo aplicou o modelo contra outros grupos de dados independentes e obteve bons resultados.

Trabalho similar foi executado por **??**), no qual 2098 dados de sequenciamento NGS foram avaliados, igualmente divididos entre alta qualidade e má qualidade. Para evitar vieses, dada a heterogeneidade dos dados, o grupo reorganizou o dataset em 3 subconjuntos. Os 47 atributos (12 atributos relacionados aos dados brutos e 35 atributos resultantes do alinhamento do sequenciamento contra as sequências alvo), foram avaliados individualmente com relação ao seu poder classificador para casa subgrupo. Neste artigo, os autores utilizaram apenas árvores de decisão para modelar os dados dos subgrupos, usando critério do índice de Gini. Foi proposto que 46 dos 47 atributos podem trazer informações sobre a qualidade do Sequenciamento. As principais limitações deste trabalho são: em alguns subgrupos

o número de dados ficou reduzido, comprometendo a confiabilidade da modelagem; a modelagem não foi submetida a uma validação externa ou com outros dados para verificar a generalidade do modelo; e o algoritmo árvores de decisão possui a desvantagem de ser sensível a pequenas mudanças nos dados de treinamento ou mudanças nos parâmetros.

??) publicaram um trabalho semelhante a **??**), porém com enfoque para controle de qualidade em variantes genéticas obtidas com a interpretação dos dados de sequenciamento NGS. Entre as metodologias utilizadas estão Florestas aleatórias, Redes neurais (função sigmoidal), AdaBoost, Classificador Quadrático, k-vizinhos mais próximos, Máquinas de vetores de suporte, Naive Bayes e Regressão Logística. Também foi verificado que Florestas aleatórias apresentou o melhor custo-benefício para classificação de variantes (F-score $>$ 0,94).

2.4 Algoritmos de Classificação

Dentro de aprendizado supervisionado, podem existir tarefas de regressão e classificação. A classificação é o processo de determinar uma categoria para um objeto a partir de um set de categorias pré-definidas. Algoritmos de classificação utilizam atributos para determinar como classificar um item em uma classe binária ou multi classes (**??**).

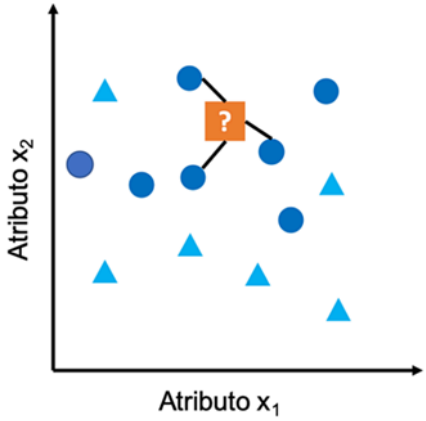
Problemas de classificação requerem exemplos rotulados (conjunto de treinamento) para determinar padrões e a acurácia do modelo é validada através de novos exemplos (conjunto de teste) que foram classificados corretamente a partir das predições (**??**). Algoritmos de classificação incluem: Árvores de Decisão, Naïve Bayes, Regressão Logística, k-vizinhos mais próximos (kNN), Máquinas de Vetores de Suporte (SVM), entre outros. A seguir serão explorados os algoritmos kNN e Árvores de Decisão.

2.4.1 k-vizinhos mais próximos

Do inglês k-nearest neighbours (kNN), este método infere a classificação de um determinado objeto com base no rótulo da maioria de seus k vizinhos mais próximos (Figura 11). Na prática, cada objeto é representado em um espaço N-dimensional (de N atributos) e os vizinhos mais próximos são determinados como

aqueles que tem a menor distância espacial em relação ao objeto cuja informação se deseja inferir (**??**). Tal distância pode ser calculada utilizando de diferentes métricas, como Euclidiana, Hamming, Manhattan, Minkowski, Ponderada, Cosseno ou Pearson. Na maioria das implementações kNN, os atributos são normalizados, para que possam ter a mesma contribuição na predição da classe (**??**).

Figura 11 – Representação gráfica do algoritmo kNN considerando 2 atributos



Fonte: **??**).

A decisão da métrica para cálculo das distâncias e o valor de 'k' devem ser ajustados experimentalmente. Geralmente altos valores de k minimizam ruídos, mas podem levar a *overfitting* (**??**).

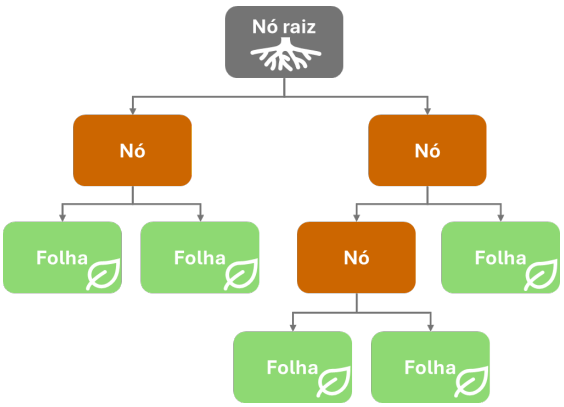
O método kNN é não paramétrico, ou seja, ele não assume premissas sobre a distribuição dos dados e não há um aprendizado do modelo por si. O método kNN apenas armazena o conjunto de dados e performa a classificação de uma nova amostra por analogia. Por isto, este modelo é considerado um modelo 'preguiçoso' (**??**).

Como vantagens, é um modelo simples de entender, eficaz e com implementação simples. Como desvantagens, é um método sensível a características irrelevantes e não é adequado para conjuntos de dados com alta dimensionalidade e/ou contendo dados faltantes.

2.4.2 Árvores de Decisão

Árvores de Decisão são fluxogramas que reduzem os dados para um conjunto de regras que podem ser usadas para uma decisão. São formadas pela seleção de variáveis para compor suas estruturas: desde a raiz, passando por cada nó interno, que representam uma decisão sobre um atributo em particular, que determinará como os dados estão particionados pelos seus nós filhos (Figura 12). Para classificar um novo exemplo, basta testar os valores dos atributos na árvore e percorrê-la até se atingir um nó folha, que irá informar o valor da classe predita (**??**).

Figura 12 – Diagrama de árvore de decisão



Fonte: elaborado pelo autor.

A divisão dos dados entre os nós da árvore pode ser baseada em diferentes métricas como índice Gini, entropia e ganho de informação. O processo de divisão é repetido múltiplas vezes até alcançar a melhor métrica de divisão dos dados ou até atingir a máxima profundidade de árvore (**??**). Dessa forma serão definidos que nós devem ir em quais posições.

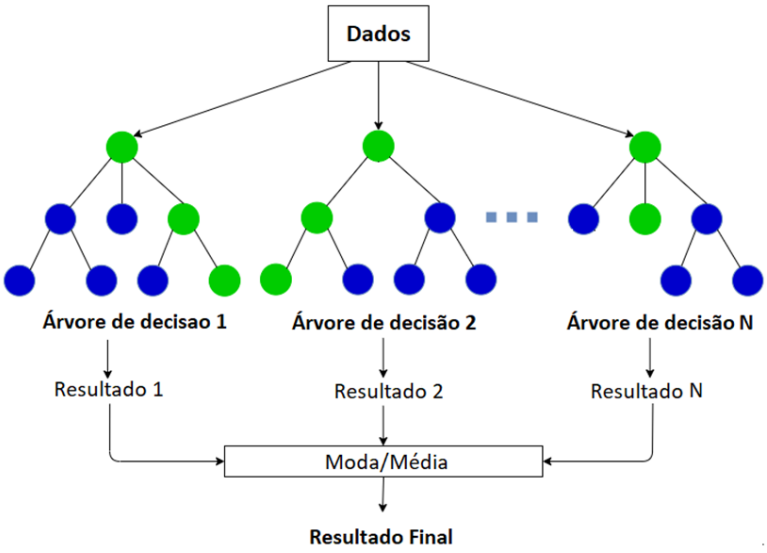
Como vantagens, árvores de decisão são simples, visuais, e por serem inspiradas na forma como humanos tomam decisão, são de melhor interpretabilidade. Possuem habilidade com dados categóricos e numéricos. Os atributos mais discriminatórios são priorizados no momento da construção da árvore, o que facilita a compreensão dos atributos com mais relevância. Como desvantagens, árvores de decisão podem levar a overfitting, principalmente se a profundidade da árvore não for definida. Para evitar esse problema, existem técnicas de 'poda' da árvore

de decisão, que tipicamente se iniciam a partir do atributo menos importante (o que provê menos informação para o modelo), ou a incorporação de técnicas de diminuição de dimensionalidade (**??**).

2.4.3 Florestas aleatórias

Florestas Aleatórias são criadas através do treino de múltiplas árvores de decisão, ao mesmo tempo e em paralelo (Figura 13). A ideia básica por trás do método é construir uma floresta de árvores de decisão individuais usando seleção aleatórias de atributos, e então obter o classificador mais frequente considerando as saídas de cada árvore de decisão (**??**). As Florestas Aleatórias utilizam, portanto, técnicas de *bagging* (treinamento paralelo) e *ensemble* (utiliza árvores de decisão individuais como aprendizes). Quanto mais árvores de decisão na floresta aleatória, mais acurados serão os resultados.

Figura 13 – Diagrama de Florestas aleatórias



Fonte: **??**)

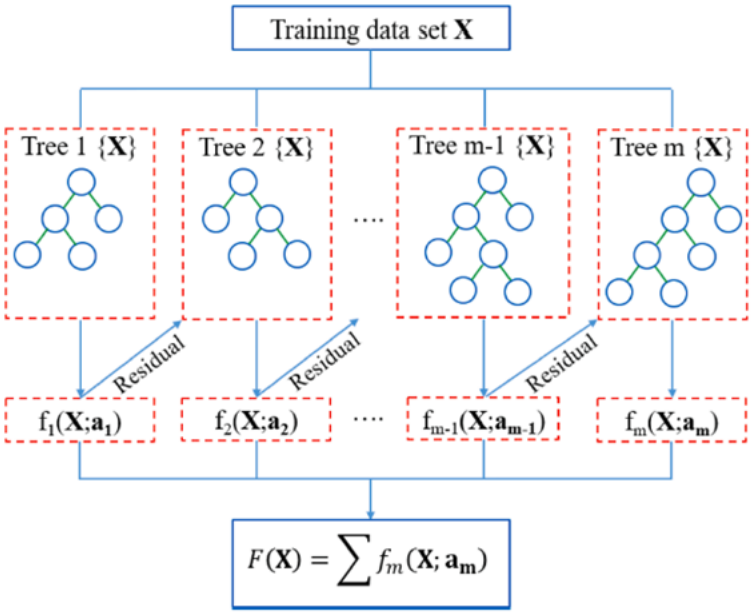
Como vantagens, as Florestas Aleatórias previnem *overfitting* e tem maior acurácia em comparação a Árvores de Decisão, funcionam bem para conjuntos de dados com alta dimensionalidade e/ou com dados faltantes. Como desvantagens,

possuem vários hiperparâmetros que precisam ser ajustados para otimização do modelo.

2.4.4 Gradient Boosting

Diferente de Florestas Aleatórias, onde árvores de decisão são geradas em paralelo por técnica de *bagging*, em algoritmos de *gradient boosting* aplicados em árvores de decisão, as árvores são treinadas sequencialmente. Cada ciclo iterativo busca por minimizar a função de perda, que ocorre pelo procedimento de gradiente descendente. O erro residual de cada árvore de decisão é levado em consideração no treinamento da próxima árvore (Figura14). Por fim é gerado um modelo final, capaz de combinar os *outputs* de todas as árvores. Entre os algoritmos de *gradient boosting*, se destacam Extreme Gradient Boosting (XGBoost) e Light Gradient Boosting machine (LightGBM).

Figura 14 – Diagrama de Gradient Boosting em árvores de decisão



Fonte: **??**).

O XGBoost usa a abordagem *level-wise tree growth*, e é um dos algoritmos de aprendizado de máquina mais utilizados. Este modelo cria histogramas de cada uma

das variáveis e os usa para encontrar a melhor partição dos dados por variável. Já o LightGBM usa a abordagem *leaf-wise tree growth* e sua principal vantagem é ganho em velocidade e uso de memória. No LightGBM são computados gradientes para cada um dos exemplos e os gradientes são usados para filtrar os dados. Exemplos com baixo gradiente já tem um bom aprendizado, enquanto que os exemplos com gradientes maiores precisam de mais aprendizado (**??**).

2.5 Considerações Finais

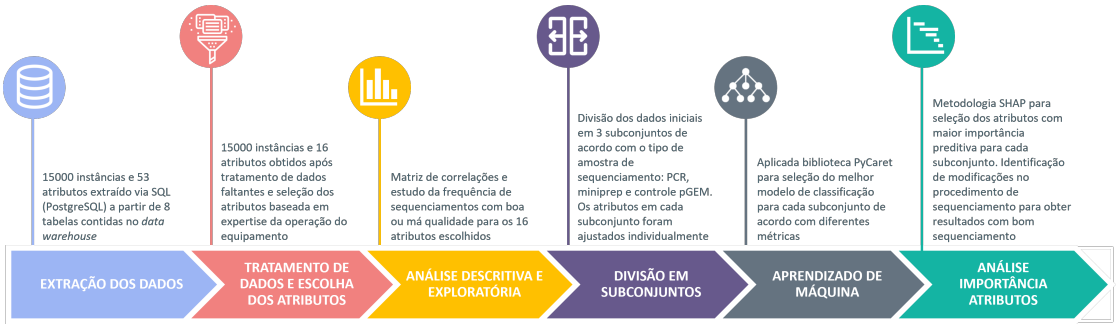
Os trabalhos apresentados anteriormente neste capitulo, focados na ciência de dados para sequenciamento, compartilham do mesmo motivador da modelagem: a automação da ação manual de avaliação de qualidade geral do sequenciamento. Naqueles trabalhos, foram encontrados excelentes valores de acurácia utilizando algoritmos classificadores, sendo que Florestas aleatórias foi o modelo mais utilizado e com melhor performance. Embora a natureza dos dados se aproxime do que será apresentado neste trabalho, naqueles foram utilizados apenas os atributos extraídos dos *trace files* para predizer a qualidade final do sequenciamento.

Este trabalho, por outro lado, visa compreender os fatores de operação do sequenciamento Sanger que contribuem para um sequenciamento de baixa ou alta qualidade. Para isso, serão considerados os atributos de operação obtidos em todas as etapas do Sequenciamento Sanger, executados nos equipamentos alocados no Laboratório de Biotecnologia da empresa "X". Através da análise e modelagem desses atributos se objetiva obter informações e esclarecimentos sobre os fatores mais relevantes durante a execução de todas as etapas deste tipo de sequenciamento, de forma a diminuir a taxa de amostras com baixa qualidade.

3 METODOLOGIA

A fim de modelar de operação de sequenciadores do tipo Sanger de um laboratório de Biotecnologia e identificar quais atributos tem maior importância na classificação de instâncias com boa ou má qualidade de sequenciamento, foram aplicados algoritmos de classificação de aprendizagem supervisionado. Entretanto, uma série de ações foram aplicadas anteriormente à modelagem, conforme representado na Figura 15 e descrito em profundidade nas seções a seguir.

Figura 15 – Etapas para modelagem dos dados de sequenciamento



Fonte: elaborado pelo autor.

3.1 Extração dos Dados

Os dados de sequenciamento do Laboratório de Biotecnologia da empresa "X"foram extraídos via SQL (PostgreSQL) a partir do *data warehouse* da empresa. 8 tabelas foram combinadas para extrair todos os dados relevantes das etapas de Preparação da amostra de DNA, PCR de terminação de cadeia, Eletroforese Capilar e Detecção e Análise dos Resultados.

Foi obtido um conjunto de dados contendo 15000 instâncias e 53 atributos iniciais, cujos dados se referem a amostras sequenciadas entre 2023 e 2024 nos sequenciadores do tipo SeqStudio (Applied Biosystems™). As amostras sequenciadas são de 3 tipos: (a) miniprep: amostra de DNA em formato circular, (b) PCR: amostra de DNA em formato linear, e (c) pGEM: amostra controle de sequenciamento

indicada pelo fabricante do sequenciador, sempre tem a mesma sequência de DNA, mesmo preparo e mesmo primer.

O conjunto de dados seguiu pelas etapas de Tratamento de dados e Escolha dos Atributos, assim sumarizada na Tabela 4. Nesta tabela também há uma breve descrição de cada um dos atributos. A análise foi implementada utilizando Python em Google Colab. As bibliotecas Numpy, Pandas, Scipy, Matplotlib e Seaborn foram utilizadas para manipulação, tratamento e visualização gráfica dos dados.

Tabela 4 – Tratamento de Dados no Conjunto de dados inicial

Atributo	**tipo atributo**	**preditora/ resposta**	**descrição**	**tratamento de dados (TD) / Escolha de atributos (EA)**
data instalação cartucho	data	preditora	data de instalação do cartucho (consumível de sequenciamento)	EA: substituído pelo novo atributo vida útil cartucho
validade cartucho	data	preditora	data de validade do cartucho	EA: substituído pelo novo atributo vida útil cartucho
data instalação do tampão	data	preditora	data de instalação do tampão (consumível de sequenciamento)	EA: substituído pelo novo atributo vida útil tampão
validade tampão	data	preditora	data de validade do tampão	EA: substituído pelo novo atributo vida útil tampão
concentracao original DNA	decimal	preditora	concentração da amostra de DNA sequenciada	EA: em redundância com atributo fmol DNA
ng DNA	decimal	preditora	massa de DNA que foi utilizada na reação da PCR de terminação de cadeia	EA: em redundância com atributo fmol DNA
tamanho fragmento	inteiro	preditora	número de bases da molécula de DNA que foi sequenciada	EA: em redundância com atributo fmol DNA
tipo amostra sequenciamento	categórica	preditora	tipo de amostra de sequenciamento: amostra controle (pGEM) ou amostra de sequenciamento (PCR, miniprep)	EA: em rendundância com atributo tipo amostra DNA
data início corrida	data	preditora	data de injeção da amostra no sequenciador	EA: em rendundância com atributo vida útil
dias de operação do cartucho	decimal	preditora	número de dias de operação do equipamento com o cartucho instalado	EA: em rendundância com atributo vida útil cartucho
dias de operação do tampão	decimal	preditora	número de dias de operação do equipamento com o tampão instalado	EA: em rendundância com atributo vida útil tampão

Tabela 4 – Tratamento de Dados no Conjunto de dados inicial

Atributo	**tipo atributo**	**preditora/ resposta**	**descrição**	**tratamento de dados (TD) / Escolha de atributos (EA)**
ab1 file	categórica	preditora	ID do arquivo ab1 gerado no sequenciamento	EA: ID
descrição amostra DNA	categórica	preditora	nome fantasia da amostra de DNA sequenciada	EA: ID
ID amostra DNA	categórica	preditora	ID da amostra de DNA sequenciada	EA: ID
placa	categórica	preditora	ID da placa de sequenciamento	EA: ID
poço	categórica	preditora	ID do poço de sequenciamento em uma matriz 8x12	EA: ID
primer	categórica	preditora	ID do primer utilizado na reação da PCR de terminação de cadeia	TD: substituição faltantes EA: ID
Classificação por cor	categórica	resposta	parâmetro de qualidade do equipamento	EA: atributo não será avaliado
CRL	decimal	resposta	parâmetro de qualidade do equipamento	EA: atributo não será avaliado
HQ%	decimal	resposta	parâmetro de qualidade do equipamento	EA: atributo não será avaliado
median pup	decimal	resposta	parâmetro de qualidade do equipamento	EA: atributo não será avaliado
RCLR	decimal	resposta	parâmetro de qualidade interno	EA: atributo não será avaliado
signal strength	decimal	resposta	parâmetro de qualidade do equipamento	EA: atributo não será avaliado
tamanho da sequência	decimal	resposta	parâmetro de qualidade do equipamento	EA: atributo não será avaliado
trace score	decimal	resposta	parâmetro de qualidade do equipamento	EA: atributo não será avaliado
data purificação	data	preditora	data da purificação da PCR de terminação de cadeia	EA: atributo com pouca relevância para modelo
data reação	data	preditora	data da reação da PCR de terminação da cadeia	EA: atributo com pouca relevância para modelo
lote bigdye	categórica	preditora	lote do reagente utilizado na reação da PCR de terminação de cadeia	TD: substituição faltantes EA: atributo com pouca relevância para modelo
lote cartucho	categórica	preditora	lote do cartucho de sequenciamento	EA: atributo com pouca relevância para modelo
lote tampão	inteiro	preditora	lote do tampão de sequenciamento	EA: atributo com pouca relevância para modelo
lote x-terminator	categórica	preditora	lote do reagente utilizado na etapa de purificação da PCR de terminação de cadeia	EA: atributo com pouca relevância para modelo

Tabela 4 – Tratamento de Dados no Conjunto de dados inicial

Atributo	**tipo atributo**	**preditora/ resposta**	**descrição**	**tratamento de dados (TD) / Escolha de atributos (EA)**
uL primer	decimal	preditora	volume de primer utilizado na reação da PCR de terminação de cadeia	EA: atributo com pouca relevância para modelo por apresentar baixíssima variação
usuário	categórica	preditora	ID do responsável pela amostra de DNA	EA: ID, atributo com pouca relevância para modelo
primer sequence	categórica	preditora	sequencia de bases do primer utilizado na reação da PCR de terminação de cadeia	TD: substituição faltantes EA: substituída por novo atributo Tm
dye set	categórica	preditora	parâmetro de operação do SeqStudio	EA: atributo fixo
Issues	decimal	preditora	frases de operação do equipamento SeqStudio	EA: atributo fixo
método bigdye	categórica	preditora	protocolo da reação da PCR de terminação de cadeia	EA: atributo fixo
método purificação	categórica	preditora	protocolo da reação de purificação	EA: atributo fixo
módulo corrida	categórica	preditora	parâmetro de operação do SeqStudio	EA: atributo fixo
tipo corrida	categórica	preditora	parâmetro de operação do SeqStudio	EA: atributo fixo
backbone*	categórica	preditora	informa detalhes do DNA circular sequenciado. Informação relevante apenas quando a amostra de DNA é do tipo miniprep	TD: substituição faltantes, one hot encoding
Capilar*	decimal	preditora	Indica qual capilar do equipamento de sequenciamento foi utilizado. O equipamento conta com 4 capilares	
DNA concentrado*	categórica	preditora	Indica se amostra foi concentrada antes do sequenciamento	TD: substituição faltantes, one hot encoding
fmol DNA*	decimal	preditora	quantidade de DNA que foi adicionada em unidade de fmol	
índice coluna*	inteiro	preditora	indica a posição vertical da amostra em uma placa de 8x12 posições	
índice linha*	categórica	preditora	Indica a posição horizontal da amostra em uma placa de 8x12 posições	TD: Transformação para variável numérica

Tabela 4 – Tratamento de Dados no Conjunto de dados inicial

Atributo	**tipo atributo**	**preditora/ resposta**	**descrição**	**tratamento de dados (TD) / Escolha de atributos (EA)**
número de injeções*	decimal	preditora	quantidade de injeções que o consumível cartucho do sequenciador executou até o momento	
protocolo purificação DNA*	categórica	preditora	método pelo qual foi purificado o DNA da amostra sequenciada	TD: one hot encoding
protocolo quantificacao DNA*	categórica	preditora	método pelo qual foi quantificado o DNA da amostra sequenciada	TD: one hot encoding
RCLR index*	categórica	resposta	parâmetro de qualidade interno. Pode assumir os valores Boa/Ruim	TD: Transformação para variável numérica
sequenciador*	categórica	preditora	Indica em qual equipamento rodou o sequenciamento propriamente dito	TD: one hot encoding
termociclador*	categórica	preditora	Indica em qual equipamento foi executada a reação de PCR de terminação de cadeia	TD: substituição faltantes, one hot encoding
tipo amostra DNA*	categórica	preditora	três tipos de DNA podem ser sequenciados: miniprep, PCR e controle pGEM	
Tm*	decimal	preditora	temperatura de anelamento do primer utilizado no sequenciamento	EA: novo atributo calculado
vida útil cartucho*	decimal	preditora	Porcentagem de vida útil do cartucho no momento da corrida de sequenciamento, considerando a data de validade do cartucho, data de instalação do cartucho e a data da corrida	EA: novo atributo calculado
vida útil tampão*	decimal	preditora	Porcentagem de vida útil do tampão no momento da corrida de sequenciamento, considerando a data de validade do tampão, data de instalação do tampão e a data da corrida	novo atributo calculado

3.2 Tratamento de dados

Do conjunto de dados inicial, foram encontrados dados faltantes nos atributos **backbone**, **DNA concentrado**, **primer, primer sequence**, **lote bigdye** e

termociclador, que foram assim tratados:

- **backbone**: 13692 dados faltantes. Para esses foi feita uma substituição dependente da variável **tipo de amostra DNA**. Para instâncias em o que **tipo de amostra DNA** era 'Miniprep', os dados faltantes foram substituídos por 'desconhecido'. Para instâncias em que **tipo de amostra DNA** era 'PCR', os dados faltantes foram substituídos por 'NA' (não se aplica);
- **DNA concentrado**: 4071 dados faltantes, substituídos pela moda do atributo;
- **primer** e **primer sequence**: 2 dados faltantes. Como as instâncias se referem a amostras controles de sequenciamento, os dados faltantes foram preenchidos de acordo com as demais amostras controle de sequenciamento;
- **lote bigdye**: 88 dados faltantes, substituído por 'desconhecido'; e
- **termociclador**: 84 dados faltantes, substituídos pela moda do atributo.

3.3 Escolha de Atributos

Considerando a alta dimensionalidade do conjunto de dados, uma primeira triagem dos atributos foi executada seguindo conhecimento de um especialista técnico em sequenciamento e considerando as seguintes premissas:

- atributos com IDs, nome de arquivos, mensagens de erro: 6 atributos foram removidos;
- atributos que não apresentaram variação de valor em todo o conjunto de dados: 6 atributos foram removidos;
- atributos que tem pouco potencial de explicar dados com bom ou mau sequenciamento: 8 atributos foram removidos;
- redundância entre atributos ou múltiplos atributos que poderiam ser transformados em um único atributo. 4 atributos foram removidos por estarem em redundância com outros atributos, e outros 7 atributos relacionados a

validade de consumíveis foram transformados para dois novos atributos (**vida útil cartucho** e **vida útil tampão**);

- seleção do atributo resposta. Ao final do sequenciamento são obtidos 9 indicadores de qualidade de sequenciamento. Inicialmente foi escolhido o indicador **RCRL index** para fazer a modelagem dos dados. Este índice foi desenvolvido no laboratório de Biotecnologia da empresa "X"como uma alternativa às métricas dadas pelo fabricante do equipamento (descritos na Tabela 2). O **RCRL index** é calculado pela porcentagem em relação ao total de bases obtida no dado de sequenciamento, do número de bases que atendem aos critérios de (a) qualidade da base de QV >=25 e (b) as duas próximas bases vizinhas devem ter qualidade média mínima de QV > 35 . Se a porcentagem de bases que atendem esses critérios for igual ou superior a 45%, o RCRL então assume o valor 'Boa'. Se abaixo de 45%, o RCRL assume o valor 'Ruim'.

Ao final da triagem, 16 atributos foram selecionados: **Backbone**, **Capilar**, **DNA concentrado**, **fmol DNA**, **índice coluna**, **índice linha**, **número de injeções**, **protocolo purificação DNA**, **protocolo quantificacao DNA**, **RCLR index (Braskem)**, **sequenciador**, **termociclador**, **tipo amostra DNA**, **Tm**, **vida útil cartucho** e **vida útil tampão**. Estes atributos estão identificados com o símbolo (*) na Tabela 4.

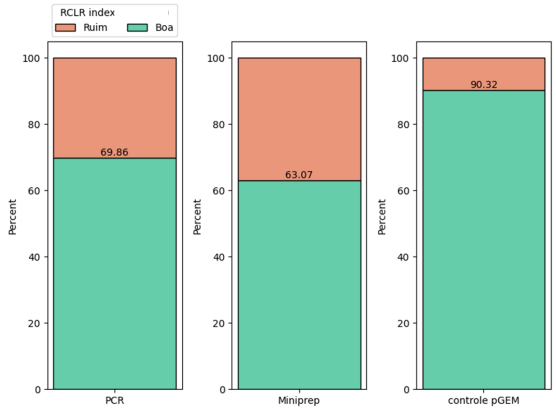
3.4 Análise Descritiva e Exploratória

Após o tratamento inicial dos dados, uma análise exploratória dos dados foi executada. Foi verificado que 69,7% dos dados de sequenciamento são classificados como boa qualidade de sequenciamento e 30,3% como baixa qualidade de sequenciamento segundo o **RCRL index**.

Os dados de classificação da qualidade de sequenciamento foram subsequentemente analisados quanto a bons ou ruins dentro dos tipos de amostra de DNA (Figura 16). Foi verificado que as amostras controle (controle pGEM) têm uma frequência de sequenciamentos com qualidade ruim abaixo de 10%. Isso está de acordo com o desejável e indicado na literatura como aceitável (Crossley *et al.*,

2020). Para as amostras do tipo PCR e Miniprep, no entanto, os valores percentis de amostras com sequenciamento ruim estão em 30,14% e 36,93%, respectivamente.

Figura 16 – Frequência relativa de amostras com sequenciamento classificado como bom ou ruim de acordo com o tipo de amostra



Ponderando o conhecimento prévio sobre o conjunto de dados e suas correlações dependendo do tipo de amostra, foi considerado que a análise exploratória seria mais adequada dividindo o conjunto de dados em 3 subconjuntos: PCR, Miniprep e controle pGEM. A divisão também foi motivada pelas diferentes frequências de tipos de amostras no conjunto de dados inicial: PCR (88,96%), miniprep (8,76%) e controle pGEM(2,28%). Como há uma prevalência grande de amostras de PCR, poderia haver um mascaramento do comportamento de determinados atributos para os outros tipos de amostras. A partir dos subconjuntos foram geradas visualizações para cada atributo individualmente, avaliando o número de instâncias do conjunto de dados classificada como Boa ou Ruim para qualidade de sequenciamento (Figuras 17 e 18). Para a maioria dos atributos, a distribuição da frequência de amostras com sequenciamento bom ou ruim parece ser homogênea entre os subconjuntos. Entretanto, alguns atributos apresentaram perfis que os destacaram dos demais:

1. **número de injeções**: predominância de sequenciamento com qualidade ruim nas primeiras injeções do cartucho (Figura 17). Pode indicar que o novo consumível instalado pode precisar de um acondicionamento antes de seu uso;

2. **Tm**: Parece haver uma predominância de sequenciamento com qualidade ruim em **Tm** maiores que 60ºC, evidente principalmente para amostras do tipo miniprep (Figura 17);

3. **fmol DNA**: há uma tendência muito interessante de concentração de amostras com boa qualidade onde ocorrem os menores valores de **fmol DNA** (Figura 17);

4. **DNA concentrado**: amostras com DNA concentrado tem maior frequência de sequenciamento com menor qualidade, especialmente para amostras de Miniprep (Figura 18). Amostras concentradas tem maior concentração de demais contaminantes que podem interferir na qualidade do resultado;

5. **capilar**: O capilar 4 parece ser o capilar com pior performance, seguido pelo capilar 1 (Figura 18);

6. **sequenciador**: há um indicativo que o sequenciador 1 está com uma maior concentração de amostras com menor qualidade de sequenciamento, especialmente para amostras do tipo Miniprep (Figura 18);

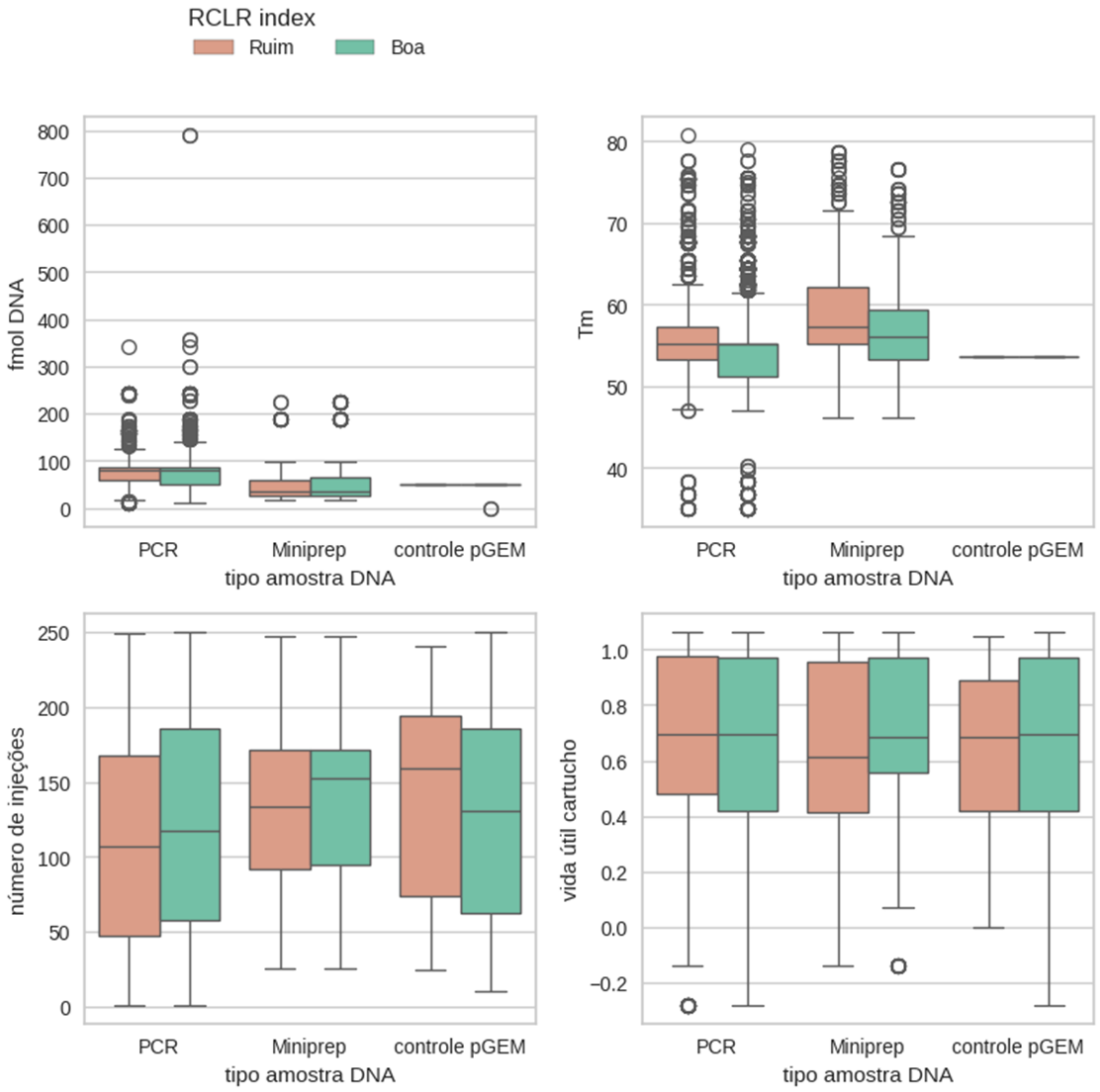
7. **protocolo purificação DNA**: amostras do tipo Miniprep com protocolo 'miniprep protocolo Cellco (adaptado para baixa cópia)' apresentam maior proporção de resultados com má qualidade (Figura 18). É esperado que nesse protocolo ocorra maior presença de contaminantes.

Figura 17 – Frequência relativa de amostras com sequenciamento classificado como bom ou ruim no conjunto inicial de dados para atributos discretos e contínuos

Figura 18 – Frequência relativa de amostras com sequenciamento classificado como bom ou ruim no conjunto inicial de dados para atributos categóricos

A temperatura de anelamento do primer, concentração de DNA, número de injeções e vida útil do cartucho representados respectivamente pelos atributos **Tm**, **fmol DNA**, **número de injeções** e **vida útil do cartucho**, foram também analisados em diagramas de caixas na Figura 19. Para o atributo **Tm** fica visível que, em média, as amostras com sequenciamento ruim foram executadas com primers com **Tm** mais alta. Para os demais atributos, as comparações não ficaram tão claras e poucas associações e correlações podem ser feitas por estas visualizações.

Figura 19 – Diagramas de caixa para seleção de variáveis contínuas



Por fim, após a codificação das variáveis categóricas, foi gerada uma matriz de correlações para cada subconjunto (Figuras 20, 21 e 22). Foi verificado que em todos os subconjuntos há uma correlação forte de valor 1,0 entre **vida útil do cartucho** e **vida útil do tampão**. Isso informa que os dois consumíveis, cartucho e tampão, sempre são instalados no mesmo dia e tem o mesmo prazo de validade após a instalação. Logo as duas variáveis **vida útil do cartucho** e **vida útil do tampão** serão resumidas a um novo único atributo **vida útil cartucho e tampão** para a posterior modelagem.

Figura 20 – Matriz de correlações para subconjunto de amostras do tipo PCR

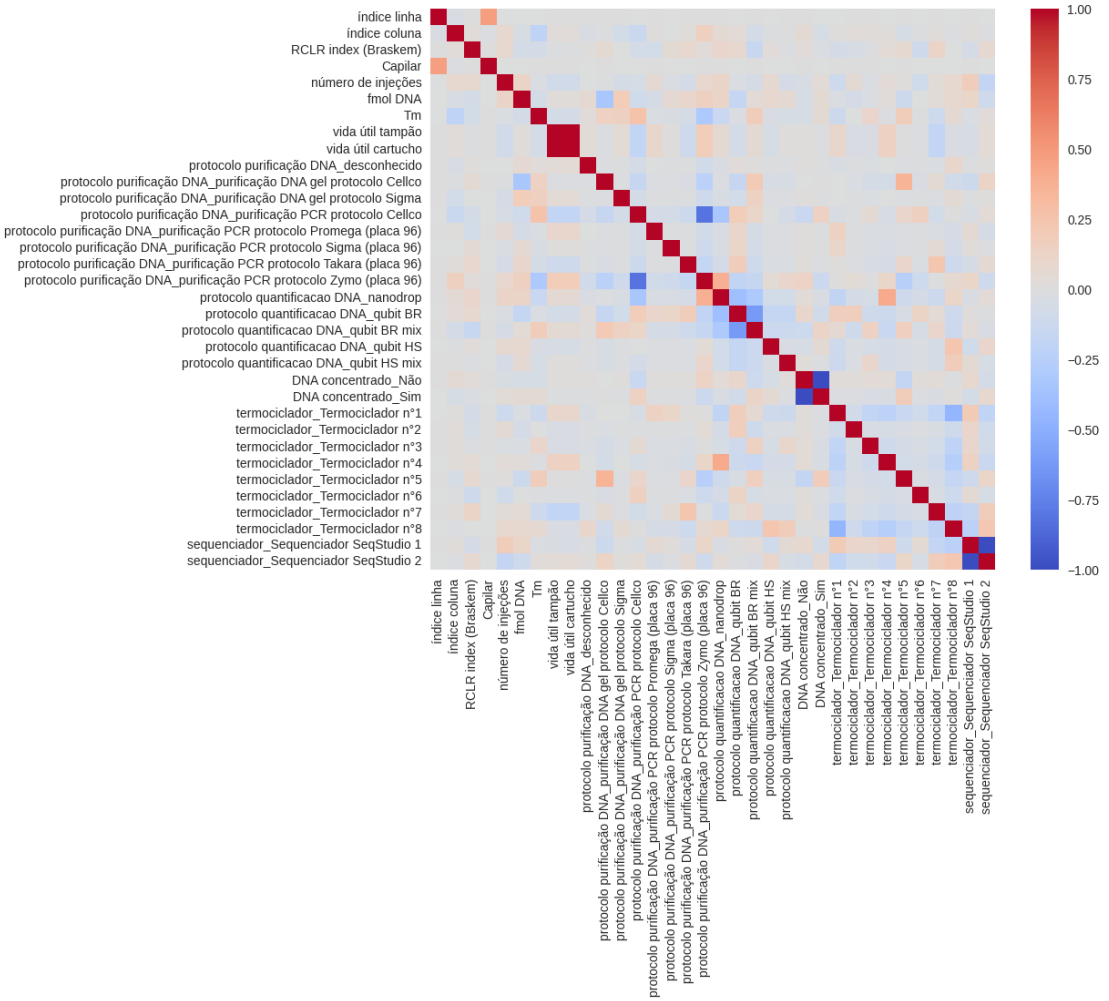


Figura 21 – Matriz de correlações para subconjunto de amostras do tipo Miniprep

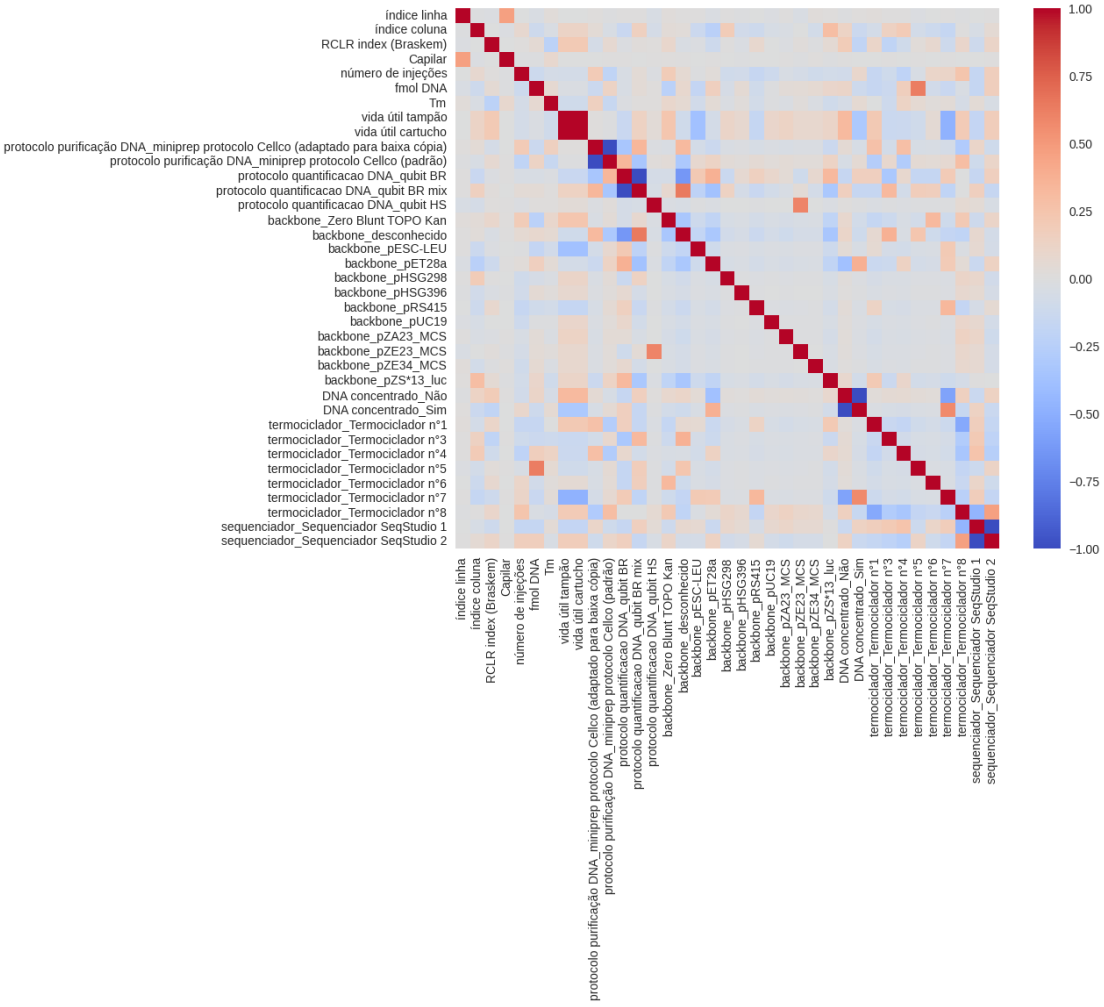
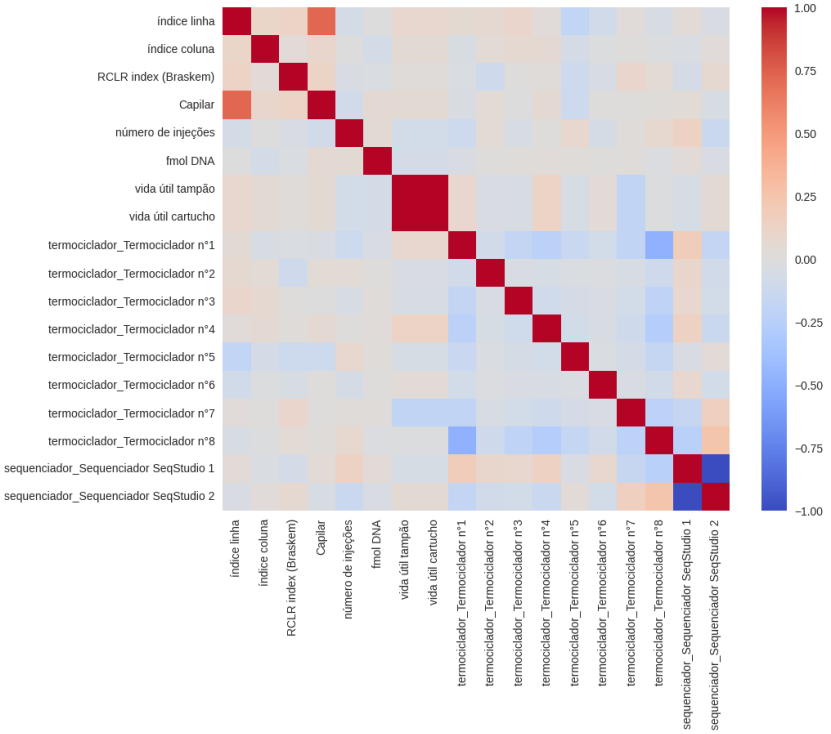


Figura 22 – Matriz de correlações para subconjunto de amostras do tipo Controle pGEM



Foi verificada uma relação positiva entre os atributos **capilar** e **índice linha**. Isso era esperado uma vez que o equipamento utilizado na no conjunto de dados tem 4 capilares que injetam simultaneamente 4 linhas de amostra por vez, sempre na mesma sequência. Portanto, amostras com índice de linha 1 e 5 sempre terão o 'capilar 1' como valor, enquanto que amostras com índice de linha 2 e 6 terão sempre o 'capilar 2' como valor e assim por diante. Apesar dos atributos **capilar** e **índice linha** serem parcialmente redundantes, ambos foram mantidos no modelo para diferenciar injeções entre as diferentes linhas. No geral, nenhum atributo preditivo apresentou uma forte correlação positiva ou negativa com o variável resposta **RCLR (index)**, o que reforça a multifatoriedade deste problema e a dificuldade em encontrar uma causa raiz para os sequenciamentos com baixa qualidade.

3.5 Aprendizado de Máquina

O PyCaret (versão 3.3.2) foi utilizado para analisar os 3 subconjuntos de dados. O PyCaret é uma biblioteca que permite a execução simultânea de algoritmos de aprendizado de máquina, de forma que todos os cálculos são embutidos e executados sequencialmente. A performance de 15 algoritmos de classificação foram comparados e analisados para cada subconjunto. O melhor algoritmo foi selecionado e os hiperparâmetros foram calibrados através da validação cruzada de 10 grupos. A performance do modelo foi avaliada usando validação cruzada estratificada de 10 grupos, SMOTE como técnica de *oversampling* para tratar o desbalanceamento dos dados e com uma divisão 70:30 para dados de treinamento:teste.

Os modelos foram avaliados com base no tempo total de treinamento e em métricas que se utilizam das quatro categorias da matriz de confusão: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). As métricas utilizadas são listadas a seguir:

Tempo total de treinamento (TT): Tempo em segundos necessário para ajustar o modelo ao conjunto de dados.

Acurácia: Mede a proporção de predições corretas em relação ao total de exemplos avaliados. É uma métrica simples e amplamente usada para modelos de classificação. A acurácia pode variar de 0 a 1 e quanto mais próximo de 1, mas acurado é um modelo. Entretanto, em problemas com classes desbalanceadas, a acurácia pode ser enganosa. A acurácia é representada pela seguinte equação:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Área sobre a Curva Característica de Operação do Receptor (AUC ROC): Avalia a capacidade do modelo de distinguir entre classes positivas e negativas, baseada na Curva ROC (Receiver Operating Characteristic), que plota a taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR). A AUC ROC pode variar entre 0 a 1 e uma AUC ROC próxima de 1 indica um modelo muito bem ajustado enquanto que uma AUC de 0,5 indica que o modelo não tem

poder preditivo e equivale a uma predição aleatória. A AUC ROC é representada pela seguinte equação:

$$\text{AUC} = \int_0^1 TPR(FPR)\, dFPR \tag{3.2}$$

Revocação: Mede a proporção de exemplos positivos corretamente identificados pelo modelo. A revocação pode variar de 0 a 1 e quanto mais próximo de 1, mas acurado é um modelo. A revocação é uma métrica muito relevante em problemas onde é crítico identificar todos os exemplos positivos. A revocação é representada pela seguinte equação:

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{3.3}$$

Precisão: Mede a proporção de exemplos positivos preditos que realmente são positivos. A precisão pode variar de 0 a 1 e quanto mais próximo de 1, mas acurado é um modelo. A precisão é uma métrica muito importante para problemas onde o custo de um falso positivo é alto. A precisão é representada pela seguinte equação:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{3.4}$$

Escore F1 (F1 Score): O escore F1 é a média harmônica da precisão e revocação, oferecendo uma métrica balanceada que considera ambas. O F1 Score pode variar de 0 a 1 e quanto mais próximo de 1, mas acurado é um modelo. O F1 é útil quando há um trade-off entre precisão e revocação e ambas são igualmente importantes. Um F1 alto indica que o modelo mantém um bom equilíbrio entre essas métricas. O F1 Score é representado pela seguinte equação:

$$F_1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.5}$$

Coeficiente Kappa: O coeficiente Kappa mede a concordância entre as predições do modelo e as classes reais, levando em consideração a concordância que ocorreria ao acaso. A métrica Kappa pode variar de -1 a 1, onde 1 indica concordância perfeita, 0 indica concordância aleatória, e valores negativos indicam desacordo. A métrica Kappa é representada pela seguinte equação:

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{3.6}$$

Coeficiente de correlação de Mattews (MCC): O MCC é uma métrica robusta que considera todas as quatro categorias da matriz de confusão e é especialmente indicada para problemas com classes desbalanceadas. O MCC pode variar de -1 a 1. Um MCC de 1 indica predição perfeita, 0 indica predição aleatória, e -1 indica total desacordo. O MCC é representado pela seguinte equação:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.7}$$

Explicabilidade Posteriormente, o melhor modelo para cada subconjunto de dados foi submetido a análise de importância dos atributos utilizando a metodologia SHAP (SHapley Additive exPlanations). Esta recentemente metodologia baseada na teoria dos jogos (Lundberg; Lee, 2017) busca trazer maior explicabilidade a modelos de aprendizado de máquina. Os valores SHAP codificam a importância que um modelo preditivo dá para um atributo, de tal forma que podemos utilizar a contribuição de cada atributo para ordená-los conforme sua importância (Marcílio; Eler, 2020). Essa metodologia permite explicar de forma gráfica e simplificada, como modelos de aprendizado de máquina chegaram em seus resultados.

4 RESULTADOS

Conforme mencionado anteriormente, três tipos de amostra podem ser corridas em um Sequenciador do tipo Sanger: PCR, miniprep e controle pGEM. Cada tipo de amostra de DNA tem suas particularidades técnicas e considerando o conjunto de dados deste trabalho, cada tipo tem também uma frequência única de dados entre as classes de qualidade de sequenciamento. Há também uma predominância expressiva de amostras do tipo PCR (cerca de 89%). Além disso, para cada tipo de amostra, um determinado grupo de atributos é aplicável. Como exemplo, o atributo **backbone** se aplica apenas para amostras do tipo miniprep, enquanto que o atributo **DNA concentrado** não tem variedade de valores para amostras do tipo pGEM. Portanto, a fim de gerar uma análise mais rica e evitar enviesamento do modelo pelo desbalanceamento do tipo da amostra, somado ao desbalanceamento das frequências de amostras com boa e má qualidade de sequenciamento, o conjunto original foi dividido em 3 subconjuntos com base no tipo de amostra. Cada qual foi modelado individualmente considerando apenas seus atributos relevantes utilizando o PyCaret para determinar qual algoritmo ajustado teve melhor desempenho na tarefa de classificação binária (qualidade de sequenciamento bom ou ruim). Os resultados serão discutidos nas seções a seguir.

4.1 Subconjunto controle pGEM

Este subconjunto de dados representa uma amostra controle, invariável, que é executada uma ou mais vezes por corrida de sequenciamento. Essa amostra tem sempre a mesma sequência de DNA, é obtido pelo mesmo protocolo de purificação, é usada sempre na mesma concentração e com o mesmo primer. Portanto, esse subconjunto tem atributos relacionados em sua grande maioria à operação do sequenciador.

Este subconjunto teve um resultado muito interessante (Tabela 5). Vários algoritmos baseados em Árvores de Decisão apresentaram acurácia acima de 0,80, mas AUC ROC entre 0,5 e 0,6. Os valores de Revocação, Escore F1 e Precisão

também estão elevados (acima de 0,90), mas os valores de Kappa e MCC estão próximos a zero. A análise das métricas indica um forte viés para a classe positiva, ou seja, o modelo não está separando bem as classes, mesmo que consiga capturar bem os positivos.

Tabela 5 – Desempenho de 15 algoritmos de aprendizado de máquina para predição de qualidade de sequenciamento para o subconjunto pGEM, ordenados em ordem decrescente de performance

Modelo	**Acurácia**	**AUC**	**Revocação**	**Precisão**	**F1**	**Kappa**	**MCC**	**TT (segundos)**
Linear Discriminant Analysis	**0,903**	**0,479**	**1,000**	**0,903**	**0,949**	**0,000**	**0,000**	**0,092**
Ridge Classifier	0,895	0,439	0,991	0,903	0,944	-0,009	-0,009	0,047
Random Forest Classifier	0,882	0,612	0,962	0,912	0,936	0,121	0,129	0,251
Gradient Boosting Classifier	0,866	0,544	0,944	0,910	0,926	0,076	0,081	0,308
Light Gradient Boosting Machine	0,866	0,497	0,944	0,911	0,926	0,096	0,098	0,417
Extreme Gradient Boosting	0,862	0,578	0,940	0,910	0,924	0,081	0,083	0,110
Extra Trees Classifier	0,857	0,566	0,939	0,906	0,922	0,030	0,033	0,193
Logistic Regression	0,841	0,452	0,926	0,901	0,912	-0,035	-0,034	0,855
Decision Tree Classifier	0,832	0,617	0,884	0,929	0,904	0,175	0,196	0,048
Ada Boost Classifier	0,832	0,591	0,898	0,916	0,905	0,092	0,097	0,168
K Neighbors Classifier	0,660	0,650	0,689	0,914	0,778	0,059	0,069	0,049
SVM - Linear Kernel	0,556	0,543	0,553	0,777	0,597	0,061	0,035	0,050
Quadratic Discriminant Analysis	0,458	0,530	0,429	0,933	0,561	0,066	0,104	0,047
Naive Bayes	0,290	0,487	0,245	0,780	0,351	-0,017	-0,046	0,048
Dummy Classifier	0,097	0,500	0,000	0,000	0,000	0,000	0,000	0,044

Legenda: as melhores métricas encontradas estão destacadas em verde.

Esse subconjunto em particular é mais desbalanceado que os demais, com menos de 10% dos dados com classificação **RCLR Index** como "Ruim"e é também o conjunto com menor quantidade de dados (341 exemplos). Pode-se concluir que, para esse subconjunto de dados, não foi possível gerar um modelo preditivo com

boa performance e que as amostras de sequenciamento com má qualidade neste subconjunto poderiam ser explicadas por atributos que não foram considerados no modelo ou então são gerados por eventos aleatórios.

4.2 Subconjunto PCR

Este subconjunto de dados representa amostras de DNA lineares de sequência desconhecida. Portanto, esse subconjunto tem atributos relacionados à operação do sequenciador e às características da amostra de DNA linear. Os resultados de performance dos algoritmos é ilustrada na Tabela 6.

Tabela 6 – Desempenho de 15 algoritmos de aprendizado de máquina para predição de qualidade de sequenciamento para o subconjunto PCR, ordenados em ordem decrescente de performance

Modelo	**Acurácia**	**AUC**	**Revocação**	**Precisão**	**F1**	**Kappa**	**MCC**	**TT (segundos)**
Extreme Gradient Boosting	**0,821**	**0,854**	**0,910**	**0,846**	**0,877**	**0,552**	**0,557**	**0,359**
Light Gradient Boosting Machine	0,818	0,857	0,917	0,838	0,876	0,539	0,547	34,500
Random Forest Classifier	0,809	0,850	0,886	0,848	0,866	0,532	0,534	18,130
Extra Trees Classifier	0,806	0,842	0,878	0,850	0,863	0,529	0,529	16,620
Gradient Boosting Classifier	0,776	0,811	0,888	0,810	0,847	0,431	0,438	22,300
Decision Tree Classifier	0,766	0,726	0,827	0,837	0,832	0,448	0,449	0,170
K Neighbors Classifier	0,757	0,811	0,774	0,863	0,816	0,459	0,465	0,264
Ada Boost Classifier	0,732	0,755	0,822	0,800	0,811	0,351	0,352	0,543
Ridge Classifier	0,666	0,677	0,705	0,794	0,747	0,262	0,266	0,223
Linear Discriminant Analysis	0,664	0,678	0,703	0,794	0,745	0,259	0,263	0,125
Logistic Regression	0,662	0,677	0,690	0,799	0,740	0,262	0,269	16,370
SVM - Linear Kernel	0,614	0,615	0,736	0,751	0,662	0,063	0,084	0,506
Naive Bayes	0,553	0,641	0,463	0,819	0,590	0,174	0,211	0,113
Quadratic Discriminant Analysis	0,301	0,500	0,000	0,000	0,000	0,000	0,000	0,129
Dummy Classifier	0,301	0,500	0,000	0,000	0,000	0,000	0,000	0,176

Legenda: as melhores métricas encontradas estão destacadas em verde.

Ao todo, 4 algoritmos baseados em Árvores de Decisão, obtiveram acurácia superior a 0,80, sendo Extreme Gradient Boosting o algoritmo com melhor performance. O modelo apresentou acurácia de 0,821, AUC ROC de 0,854 e tempo de treinamento de 0,359 segundos, uma notória qualidade e uma das principais vantagens desse algoritmo em relação às Florestas Aleatórias. Além disso, obteve valores de Revocação de 0,910, Precisão 0,846 e Escore F1 de 0,877. Os coeficientes Kappa e MCC foram de 0,552 e 0,557, respectivamente. Os gráficos das principais métricas são ilustrados na Figura 23.

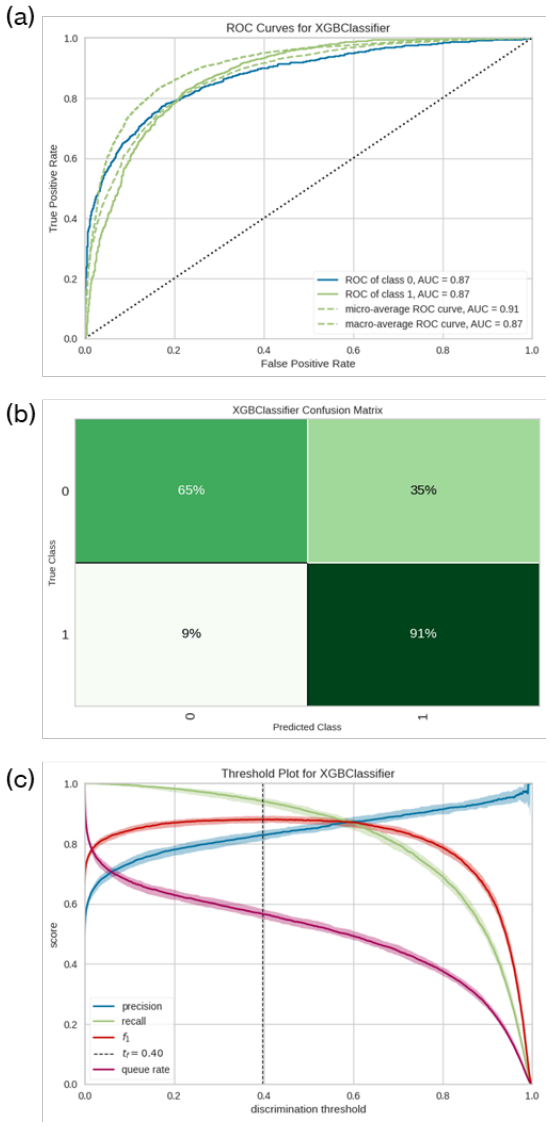
Outro ponto importante a ser considerado é que o algoritmo kNN, que possui lógica muito diferente ao conceito de Florestas Aleatórias e suas variações (LightGBM e XGBoost), obteve um resultado satisfatório com acurácia de 0,757 e AUC ROC de 0,811. Este algoritmo é de fácil implementação e entendimento e poderia ser considerado como uma opção para aprendizado de máquina dos dados descritos neste trabalho.

O modelo XGBoost foi tunado utilizando o PyCaret. Entretanto, as métricas de desempenho não foram melhores que o modelo original. Portanto, as análises subsequentes serão feitas considerando o modelo original.

A curva ROC é uma representação visual do desempenho do modelo em todos os limites, desenhada pelo cálculo da taxa de verdadeiro positivo e a taxa de falsos positivos em intervalos selecionados. No subconjunto PCR podemos ver a ótima performance do modelo em discriminar os verdadeiros positivos, com AUC ROC de 0,87 (Figura 23a). Na matriz de confusão (Figura 23b), se pode confirmar também a capacidade do modelo em diferenciar verdadeiros positivos de falsos negativos, com apenas 9% de exemplos positivos sendo erroneamente classificados como negativos. Entretanto é possível observar que o modelo encontra mais dificuldade na classificação dos exemplos negativos, com acerto de 65% dos exemplos.

O Gráfico de Limiar de Discriminação permite visualizar a relação entre 4 métricas de performance e o limiar (*threshold*) de discriminação. A linha pontilhada indica o limiar discriminatório ótimo, relativo ao ponto máximo da métrica F1 Score. Conforme observado na Figura 23c, o limiar discriminatório ótimo do modelo é próximo a 0,4. Neste limiar, o valor de *queue rate* (que representa a porcentagem de todos os positivos preditos sobre todos os exemplos) é de 0,57. Ou seja, o modelo marca os exemplos com pontuação presente nos 57 percentis superiores como positiva.

Figura 23 – Métricas de performance para modelo de classificação XGBoost aplicado para subconjunto de dados PCR



Legenda: (a) Curva AUC, (b) Matriz de Confusão e (c) Gráfico de Limiar de Discriminação.

4.3 Subconjunto Miniprep

Este subconjunto de dados representa amostras de DNA circulares de sequência desconhecida. Portanto, esse subconjunto tem atributos relacionados

à operação do sequenciador e à característica de obtenção da amostra de DNA circular. Os resultados de performance dos algoritmos é ilustrada na Tabela 7.

Tabela 7 – Desempenho de 15 algoritmos de aprendizado de máquina para predição de qualidade de sequenciamento para o subconjunto miniprep, ordenados em ordem decrescente de performance

Modelo	**Acurácia**	**AUC**	**Revocação**	**Precisão**	**F1**	**Kappa**	**MCC**	**TT (segundos)**
Light Gradient Boosting Machine	**0,751**	**0,792**	**0,825**	**0,790**	**0,806**	**0,457**	**0,460**	**0,323**
Extreme Gradient Boosting	0,745	0,789	0,816	0,788	0,801	0,447	0,449	0,194
Random Forest Classifier	0,741	0,795	0,815	0,784	0,798	0,437	0,440	0,402
Extra Trees Classifier	0,731	0,784	0,815	0,772	0,792	0,412	0,415	0,508
Gradient Boosting Classifier	0,729	0,761	0,827	0,765	0,794	0,400	0,404	0,279
K Neighbors Classifier	0,697	0,745	0,712	0,789	0,747	0,372	0,377	0,064
Decision Tree Classifier	0,696	0,677	0,751	0,763	0,756	0,353	0,354	0,078
Ada Boost Classifier	0,674	0,717	0,745	0,740	0,742	0,299	0,300	0,181
Linear Discriminant Analysis	0,671	0,687	0,733	0,745	0,738	0,296	0,298	0,055
Ridge Classifier	0,669	0,691	0,726	0,746	0,735	0,294	0,295	0,096
Logistic Regression	0,667	0,691	0,721	0,745	0,732	0,291	0,293	0,698
Naive Bayes	0,628	0,666	0,700	0,718	0,688	0,206	0,217	0,057
SVM - Linear Kernel	0,516	0,636	0,421	0,625	0,428	0,100	0,118	0,108
Dummy Classifier	0,369	0,500	0,000	0,000	0,000	0,000	0,000	0,049
Quadratic Discriminant Analysis	0,366	0,463	0,002	0,033	0,003	-0,008	-0,039	0,063

Legenda: as melhores métricas encontradas estão destacadas em verde.

Da mesma forma que observado para o subconjunto PCR, os algoritmos baseados em árvores de decisão foram os que obtiveram acurácia superior entre os modelos avaliados, e Light Gradient Boosting foi o algoritmo com melhor performance. O modelo apresentou acurácia de 0,751, AUC ROC de 0,792 e tempo de treinamento de 0,323 segundos. Além disso, obteve valores de Revocação de

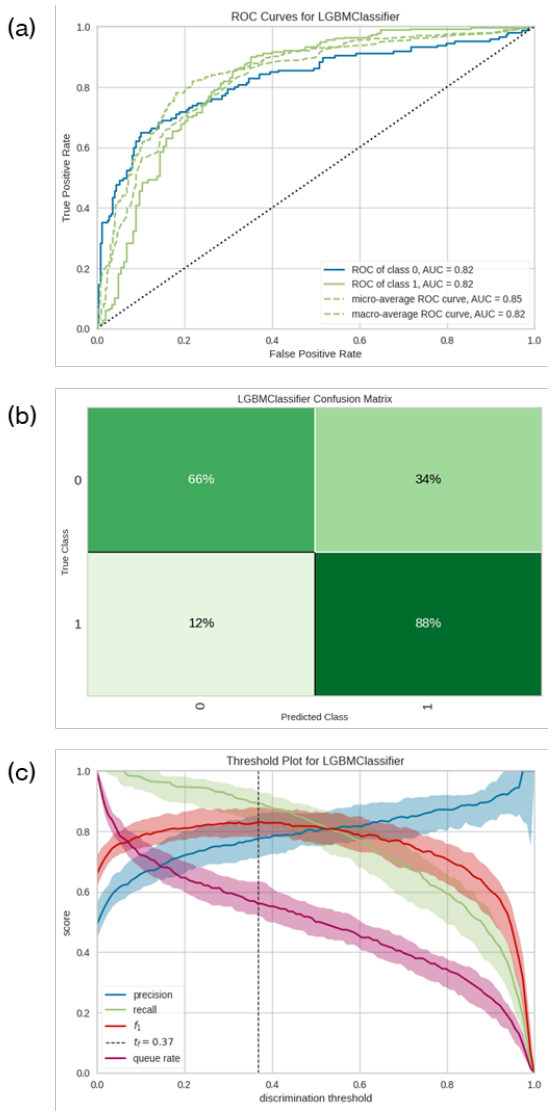
0,825, Precisão 0,790 e Escore F1 de 0,806. Os coeficientes Kappa e MCC foram de 0,457 e 0,460, respectivamente. Os gráficos das principais métricas são ilustrados na Figura 24.

O modelo LightGBM foi tunado utilizando o PyCaret. Entretanto, da mesma forma que observado para o subconjunto PCR, as métricas de desempenho não foram melhores que o modelo original. Portanto, as análises subsequentes serão feitas considerando o modelo original.

No subconjunto miniprep podemos ver a ótima performance do modelo em discriminar os verdadeiros positivos, com AUC ROC de 0,82 (Figura 24a). Da mesma forma que observado para o subconjunto PCR, o modelo é capaz de diferenciar verdadeiros positivos de falsos negativos, com apenas 12% de exemplos positivos sendo erroneamente classificados como negativos, evidenciado na matriz de confusão (Figura 24b). O modelo encontra mais dificuldade na classificação dos exemplos negativos, com acerto de 66% dos exemplos.

Conforme observado no Gráfico de Limiar de Discriminação na Figura 24c, o limiar discriminatório ótimo do modelo é próximo a 0,37. Neste limiar, o valor de *queue rate* é também próximo a 0,57.

Figura 24 – Métricas de performance para modelo de classificação LightGBM aplicado para subconjunto de dados miniprep



Legenda: (a) Curva AUC, (b) Matriz de Confusão e (c) Gráfico de Limiar de Discriminação.

4.4 Análise de Importância de Atributos

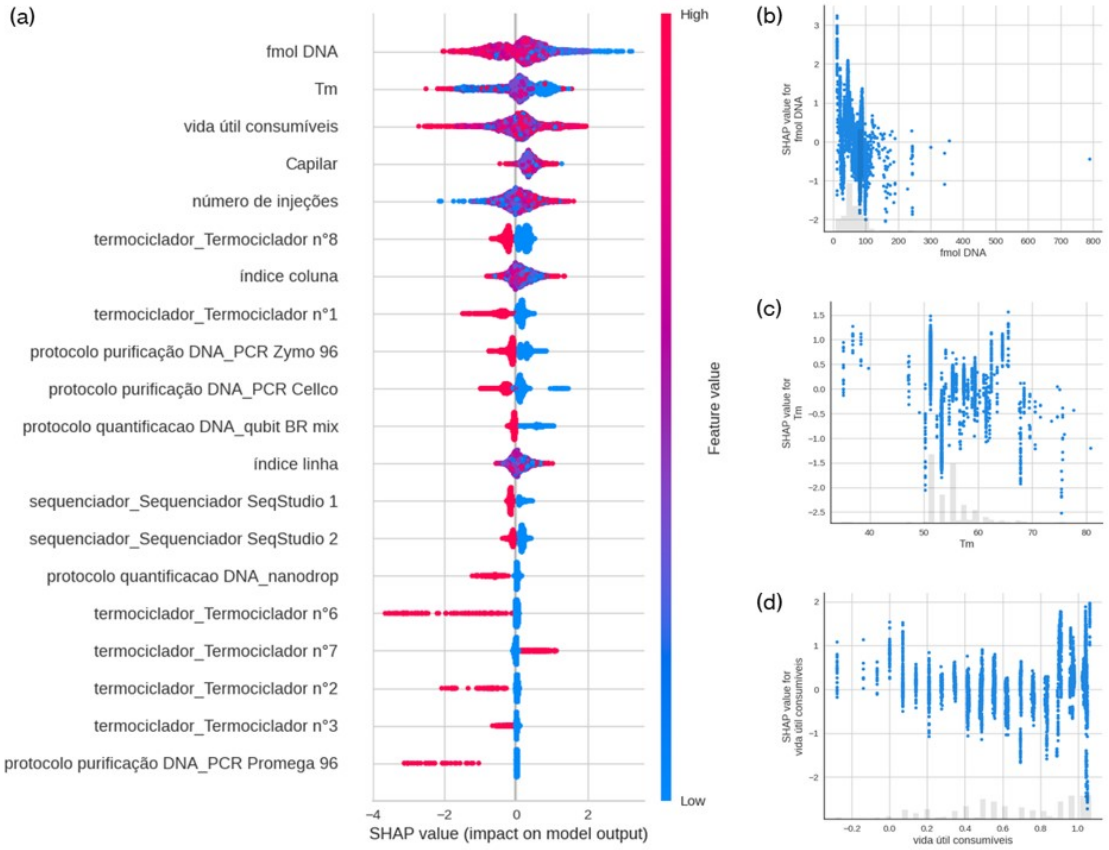
A principal questão no estudo de qualidade de dados de sequenciamento envolve o entendimento de quais atributos tem impacto sobre a qualidade do

sequenciamento e como traduzir essa informação em mudanças no protocolo desde a obtenção das amostras de DNA purificadas até a operação do sequenciador, de forma a maximizar amostras de DNA sequenciadas com boa qualidade.

Para listar o impacto dos atributos mais importantes para cada modelo e seu subconjunto de amostras, foi utilizado o gráfico SHAP. Os valores SHAP fornecem uma interpretação do impacto de cada atributo na previsão do modelo. Quanto maior o valor SHAP de um atributo, maior o impacto desse atributo na previsão final do modelo.

Para amostras do tipo PCR, os 3 atributos mais informativos de forma global são **fmol DNA**, **Tm** e **vida útil consumíveis** (Figura 25). Para amostras do tipo miniprep, os 3 atributos com maior valor informativo são **Tm**, **vida útil consumíveis** e **numero de injeções** (Figura 26).

Figura 25 – Gráficos SHAP para investigação de importância de atributos no modelo XGBoost para subconjunto PCR

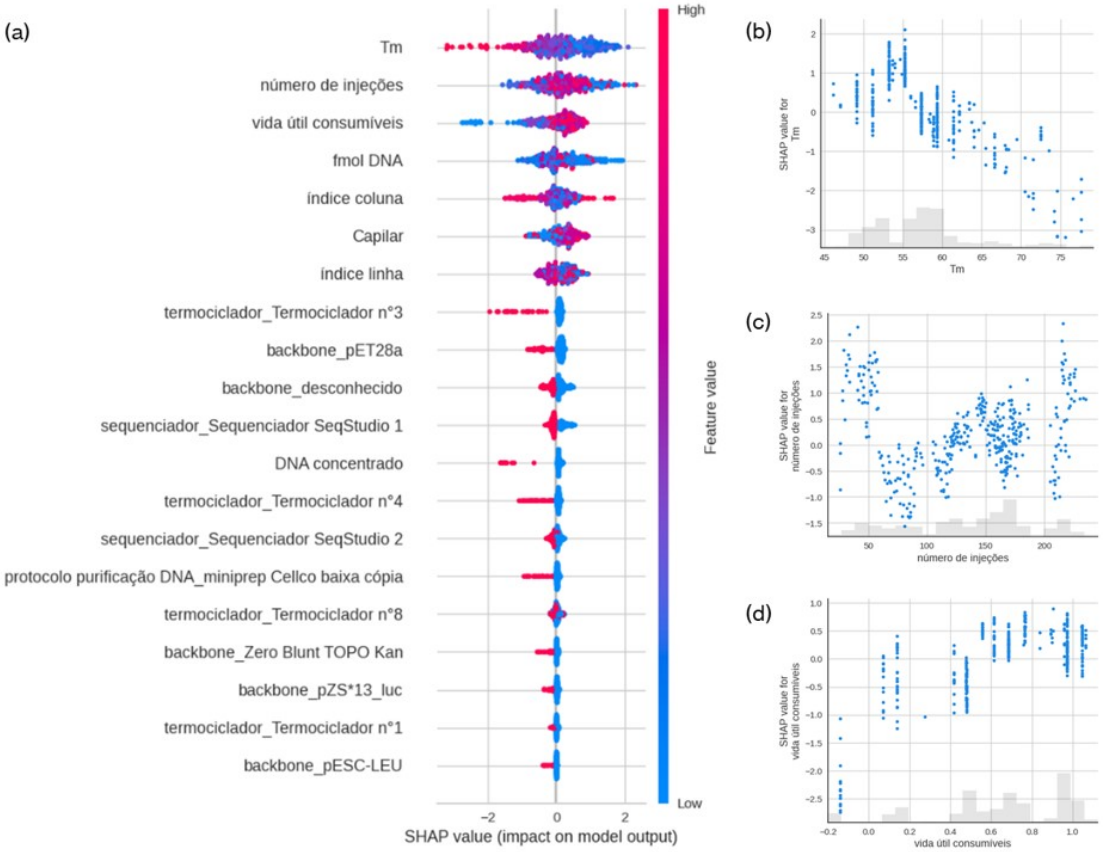


Legenda: (a) visão geral do rankeamento de atributos, com atributos de maior importância no topo do gráfico e (b), (c) e (d) gráficos de dependência dos 3 atributos com maior importância no modelo.

Nas amostras do tipo PCR (Figura 25), pode-se observar que o atributo **fmol DNA** tem uma contribuição negativa para as predições. Quanto maior o valor de fmol DNA (mais vermelho), maior é a tendência do valor SHAP ser mais negativo, contribuindo então para um valor de predição da classe negativa, o que corresponde a **RCLR index** de má qualidade. O atributo **Tm** tem um comportamento similar: quanto maior o valor deste atributo, maior é a tendência do valor SHAP ser negativo. O atributo **vida útil dos consumíveis**, no entanto tem uma dispersão maior dos dados próximo ao valor 1, o que dificulta a interpretação

da dependência do valor SHAP com o valor do atributo. O valor de 1,0 neste atributo é referente ao consumível recém instalado no equipamento.

Figura 26 – Gráficos SHAP para investigação de importância de atributos no modelo LightGBM para subconjunto miniprep



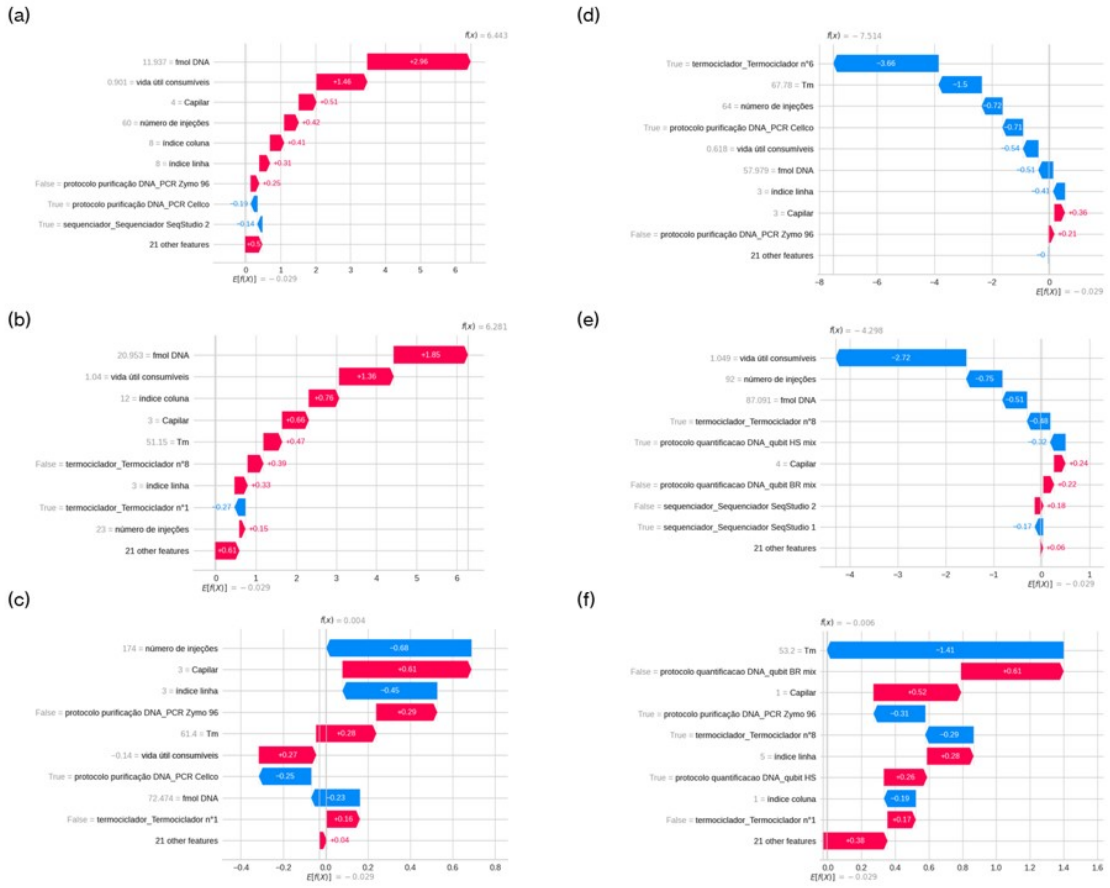
Legenda: (a) visão geral do rankeamento de atributos, com atributos de maior importância no topo do gráfico e (b), (c) e (d) gráficos de dependência dos 3 atributos com maior importância no modelo.

Para amostras do tipo miniprep (Figura 26), pode-se observar que o atributo **Tm** tem uma contribuição negativa para as predições. Quanto maior o valor de **Tm** (mais vermelho), maior é a tendência do valor SHAP ser mais negativo, contribuindo então para um valor de predição da classe negativa. O atributo **vida útil dos consumíveis**, tem o comportamento inverso: quanto menor, maior a tendência do valor SHAP ser negativo. Ou seja, quando os consumíveis estão chegando ao fim

de sua vida útil (próximos a 0) há uma tendência em obter resultados com menor qualidade. O atributo **número de injeções** tem um comportamento não linear.

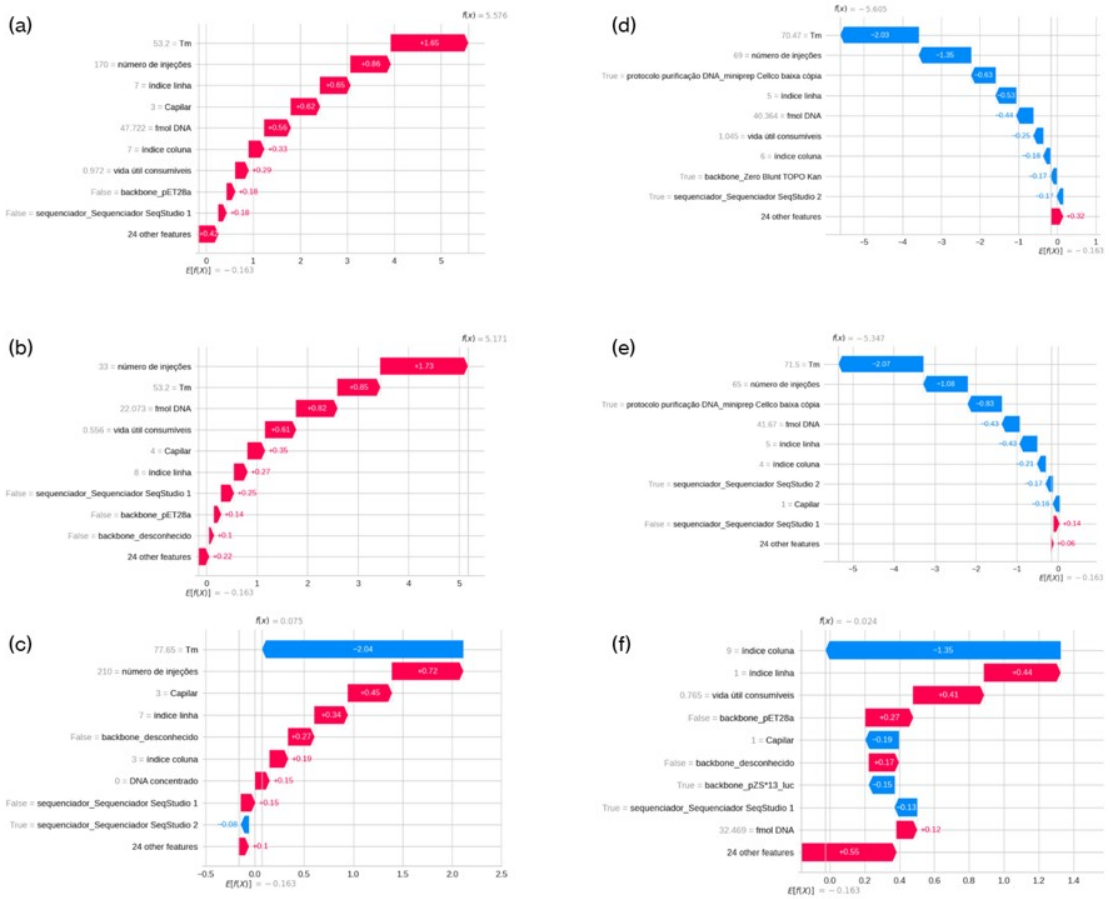
A grande vantagem do uso do SHAP está no entendimento como cada atributo contribuiu para a predição de cada exemplo individualmente. Essa propriedade será utilizada para entendimento dos falsos negativos nos subconjuntos PCR e miniprep (Figuras 27 e 28).

Figura 27 – Distribuição de valores SHAP dos principais atributos na predição de classe para 6 exemplos do subconjunto PCR utilizando o modelo XGBoost



Legenda: (a) exemplo de verdadeiro positivo, (b) exemplo de falso positivo com alta nota de predição, (c) exemplo de falso positivo com baixa nota de predição. (d) exemplo de verdadeiro negativo, (e) exemplo de falso negativo com alta nota de predição, (f) exemplo de falso negativo com baixa nota de predição.

Figura 28 – Distribuição de valores SHAP dos principais atributos na predição de classe para 6 exemplos do subconjunto miniprep utilizando o modelo LightGBM



Legenda: (a) exemplo de verdadeiro positivo, (b) exemplo de falso positivo com alta nota de predição, (c) exemplo de falso positivo com baixa nota de predição. (d) exemplo de verdadeiro negativo, (e) exemplo de falso negativo com alta nota de predição, (f) exemplo de falso negativo com baixa nota de predição.

Analisando um exemplo de falso negativo para o subconjunto PCR (Figura 27f), é possível verificar que, diferente da análise global, os dois atributos com maior importância no modelo de forma global (**fmol DNA** e **vida útil consumíveis**) não aparecem entre os 9 atributos mais importantes na predição deste exemplo em específico. Além disso, o valor de saída do exemplo (-0,006) tem um valor muito próximo ao valor de *cut-off* (-0,029). Ou seja, o exemplo está próximo da

classificação correta (verdadeiro positivo).

Outro exemplo de falso negativo para o subconjunto PCR é ilustrado na Figura 27e. Nesse caso, ao contrário do exemplo anterior de falso negativo, existe a presença de **fmol DNA** e **vida útil consumíveis** como atributos relevantes para predição do exemplo e o valor de saída é de -4,298, bem distante do *cut-off*. Uma hipótese para esse caso é que tenha ocorrido erro aleatório de sequenciamento.

Análise semelhante pode ser feita para o subconjunto miniprep. No exemplo de falso negativo da Figura 28f, é possível verificar que os dois atributos com maior importância no modelo de forma global (**Tm** e **número de injeções**) não são listados entre os 9 atributos mais importantes na predição deste exemplo. O valor de saída do exemplo (-0,024) tem valor muito próximo ao valor de *cut-off* (-0,163).

No segundo exemplo de falso negativo para o subconjunto miniprep (Figura 28e), **Tm** e **número de injeções** estão presentes como atributos relevantes para predição do exemplo e o valor de saída é bem distante do *cut-off* (-5,347). Novamente, a hipótese para esse caso é que tenha ocorrido erro aleatório de sequenciamento.

4.5 Considerações Finais

Foi possível modelar os dados de sequenciamento do tipo Sanger com relação à classificação da qualidade de amostras. Os algoritmos com melhor desempenho foram XGBoost e LightGBM para amostras de DNA linear (PCR) e circular (miniprep), respectivamente. Os modelos apresentaram poder preditivo satisfatório e forneceram *insights* sobre os atributos mais relevantes para a tarefa de classificação binária. Com essas informações, são sugeridas as seguintes medidas (em ordem de prioridade) para diminuir a quantidade de sequenciamento com baixa qualidade:

- operar em menores faixas de **fmol DNA** para amostras do tipo PCR;
- operar em **Tm** menores (até 60ºC) para amostras do tipo miniprep;
- fazer acondicionamento dos consumíveis, evitando injeções logo após sua instalação (**número de injeções**); e
- evitar injeções próximo ao fim da **vida útil dos consumíveis**.

5 CONCLUSÃO

A compreensão dos fatores de operação do sequenciamento Sanger que contribuem para um sequenciamento de baixa ou alta qualidade é uma informação essencial para diminuir a taxa de sequenciamentos com baixa qualidade e colaborar para a efetividade de operação de laboratórios que se apoiam nessa técnica para condução de projetos de Biotecnologia. Através de uma seleção técnica inicial de atributos, foi possível modelar um conjunto de dados de sequenciamento do tipo Sanger utilizando algoritmos de classificação binários baseados em árvores de decisão. Os tipos de amostras de sequenciamento foram separadas entre 3 diferentes subconjuntos de dados para o treinamento: PCR, miniprep e controle pGEM.

Amostras do tipo controle pGEM geraram modelo com baixo poder preditivo para amostras com má qualidade de sequenciamento, indicando que para esse subconjunto, as amostras com má qualidade de sequenciamento são provavelmente aleatórias e não podem explicadas pelos atributos selecionados. Para amostras do tipo PCR, foi obtido um modelo baseado em XGBoost e para amostras do tipo miniprep, foi obtido um modelo baseado em LightGBM, ambos baseados em Árvores de Decisão, com aplicação da técnica de *Gradient Boosting* para otimização do aprendizado.

Os atributos foram comparados quanto à sua importância para a predição dos modelos usando valores SHAP e foi constatado que os atributos **fmol DNA**, **Tm**, **vida útil consumíveis** e **numero de injeções** são os mais importantes nas modelagens. Esses atributos são relacionados às etapas PCR de terminação de cadeia e Eletroforese Capilar do sequenciamento. Considerando os resultados, foi possível sugerir alterações no protocolo de sequenciamento para aumentar a chance de sequenciamento com boa qualidade de sequenciamento. É importante destacar que, até a presente momento, não temos conhecimento de referências na literatura de trabalhos de aprendizado de máquina focados na predição de qualidade de sequenciamento.

Como trabalhos futuros, há várias opções para serem exploradas. Os dados

utilizados nesse trabalho não são balanceados, sendo predominante os dados com boa qualidade de sequenciamento. Como sugestão, outras técnicas de balanceamento poderiam ser aplicadas para melhorar o poder preditivo do modelo para exemplos com qualidade de sequenciamento ruim. Além disso, antes da modelagem houve uma seleção técnica de atributos. Uma opção seria obter nova modelagem incluindo os atributos anteriormente excluídos ou uma nova seleção destes. Outra possibilidade relacionada seria substituir a seleção técnica de atributos por uma estratégia de redução de dimensionalidade dos atributos iniciais através da análise do tipo Componente Principal. Uma terceira opção seria explorar a modelagem utilizando outros atributos de resposta. Neste trabalho, em específico, foi utilizada uma métrica interna da empresa "X"como classificador, porém outras métricas fornecidas pelo sequenciador poderiam ser também avaliadas, aumentando o alcance de aplicabilidade deste trabalho para mais laboratórios de Biotecnologia.

REFERÊNCIAS

CARBONELL, P. *et al.* An automated design-build-test-learn pipeline for enhanced microbial production of fine chemicals. **Communications Biology**, v. 1, n. 66, p. 1–10, 2018.

CROSSLEY, B. *et al.* Guidelines for sanger sequencing and molecular assay monitoring. **Communications Biology**, v. 32, n. 6, p. 767–775, 2020.

HEATHER, J. M.; CHAIN, B. The sequence of sequencers: The history of sequencing dna. **Genomics**, v. 107, n. 1, p. 1–8, 2016.

Life Technologies. **Troubleshooting Sanger sequencing data, MAN0014435. Revision A.0. [S. l.]**. Thermo Fisher Scientific Inc., 2016. Disponível em: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0014435_Trbleshoot_Sanger_seq_data_UB.pdf. Acesso em: 23 apr. 2024.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *In*: GUYON, I. *et al.* (ed.). **Advances in Neural Information Processing Systems**. [*S.l.: s.n.*]: Curran Associates, Inc., 2017. v. 30.

MARCíLIO, W. E.; ELER, D. M. From explanations to feature selection: assessing shap values as feature selection mechanism. *In*: **2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [*S.l.: s.n.*], 2020. p. 340–347.

NIELSEN, J.; KEASLING, J. D. Engineering cellular metabolism. **Cell**, v. 164, p. 1185–1197, 2016.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 12, p. 5463–5467, 1977.