UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Gabriel Augusto Prevato

Implementação de técnicas de aprendizado de máquina e reconhecimento de padrões para prospecção de exosítios de proteínas como moduladores de interação proteína-DNA e proteína-RNA

São José do Rio Preto
2024

Gabriel Augusto Prevato

Implementação de técnicas de aprendizado de máquina e reconhecimento de padrões para prospecção de exosítios de proteínas como moduladores de interação proteína-DNA e proteína-RNA

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: FAPESP - Proc. número 2023/13610-3

Orientador:

Prof. Dr. Geraldo Francisco Donegá Zafalon

São José do Rio Preto
2024

P944i

Prevato, Gabriel Augusto

Implementação de técnicas de aprendizado de máquina e reconhecimento de padrões para prospecção de exosítios de proteínas como moduladores de interação proteína-DNA e proteína-RNA / Gabriel Augusto Prevato. -- São José do Rio Preto, 2024

49 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Geraldo Francisco Donegá Zafalon

1. Ciência da computação. 2. Bioinformática. 3. Aprendizado do computador. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Gabriel Augusto Prevato

Implementação de técnicas de aprendizado de máquina e reconhecimento de padrões para prospecção de exosítios de proteínas como moduladores de interação proteína-DNA e proteína-RNA

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Financiadora: FAPESP

Comissão Examinadora:

Prof. Dr. Geraldo Francisco Donegá Zafalon
UNESP – Câmpus de São José do Rio Preto
Orientador

Prof. Dr. Arnaldo Candido Junior
UNESP – Câmpus de São José do Rio Preto

Prof. Dra. Carina Alexandra Rondini
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
2024

“They can take your world, they can take your heart cut you loose from all you know. But if it’s your fate then every step forward,will always be a step closer to home.”

- Sora (Kingdom Hearts III)

Agradecimentos

Agradeço aos meus pais, Penha e João, e às minhas irmãs, Giovana e Josiane, por todo apoio e esforço que tiveram para que eu pudesse seguir meus sonhos e continuar estudando, sempre me incentivando e ajudando nos momentos mais difíceis dessa jornada.

Agradeço aos meus orientadores, Geraldo e Vitória, e colegas de laboratório pelos conselhos, ajudas e conversas que tivemos durante nossas reuniões.

Agradeço a todos os meus amigos, especialmente ao Douglas, Daphne, Bruno, Gustavo, Luiza, Lucas e Gab, que foram pessoas incríveis comigo desde o momento em que as conheci, e que me ajudaram a estar em um lugar onde me senti acolhido e parte de um grupo. Obrigado por todas as noites que passamos jogando algo ou indo a uma lanchonete, pelos risos e pelos momentos que passamos juntos.

Por fim, agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro a este trabalho.

Resumo

Interações proteína-DNA e proteína-RNA desempenham papéis muito importantes na manutenção da vida, tornando-se, assim, alvos de estudos que buscam compreender mais profundamente o funcionamento celular, a fim de possibilitar o desenvolvimento de novos medicamentos e vacinas. Visando expandir o conhecimento acerca destas ligações, estudos sobre a identificação dos sítios de ligação proteína-DNA/RNA têm sido realizados tradicionalmente por meio do uso de métodos experimentais, as quais geram altos custos e uma alta demanda de tempo. Deste modo, este trabalho tem como objetivo utilizar métodos computacionais, como aprendizado de máquina e *ensemble learning*, para o desenvolvimento de um classificador capaz de realizar predições dos exosítios que modulam as ligações proteína-DNA/RNA, diminuindo assim os custos e o tempo demandados em comparação aos métodos experimentais.

Palavras-chave: Interações proteína-DNA/RNA, métodos computacionais, aprendizado de máquina, *ensemble learning*, predição de exosítios.

Abstract

Protein-DNA and protein-RNA interactions play very important roles in maintaining life, thus becoming targets of studies that seek to deepen the understanding of cellular functioning, aiming to enable the development of new drugs and vaccines. In order to expand the knowledge about these interactions, studies on the identification of protein-DNA/RNA binding sites have traditionally been conducted using experimental methods, which generate high costs and demand a significant amount of time. Therefore, this work aims to use computational methods, such as machine learning and ensemble learning, to develop a classifier capable of predicting the exosites that modulate protein-DNA/RNA interactions, thus reducing the costs and time demanded compared to experimental methods.

Keywords: Protein-DNA/RNA interactions, computational methods, machine learning, ensemble learning, exosites prediction.

Lista de Figuras

Lista de Tabelas

Sumário

Lista de Abreviações

ANN - *Artificial neural networks*

AUC - *Area under the ROC curve*

AUPRC - *Area under the precision-recall curve*

AUROC - *Area under the receiver operating characteristic*s

CNN - *Convolutional Neural Network*

DNA - Ácido desoxirribonucleico

GBDT - *Gradient Boosted Decision Tree*

IA - Inteligência Artificial

MCC - *Matthews correlation coefficient*

RFE - *Recursive Feature Elimination*

RNA - Ácido ribonucleico

RNN - *Recurrent Neural Network*

SFS - *Sequential Feature Selection*

SVM - *Support Vector Machine*

Capítulo 1

Introdução ao trabalho

1.1 Introdução

Os exosítios são regiões da superfície das proteínas que atuam como sítios de ligação secundários, exercendo um papel importante para interação entre proteínas e ligantes (BOCK; PANIZZI; VERHAMME, 2007; OBAHA; NOVINEC, 2023). Uma de suas funções é atuar como moduladores nas ligações proteína-DNA e proteína-RNA, sendo esse tipo de interação parte de processos como replicação gênica e tradução, de suma importância para as atividades celulares (CUI et al., 2022).

Um dos possíveis avanços trazidos pelos estudos nesta área é o desenvolvimento de fármacos, pois ao compreender as interações entre proteínas e ácidos nucleicos (DNA e RNA), é possível entender o funcionamento celular e desta forma facilitar o processo de criação de novos medicamentos e vacinas (CUI et al., 2022).

Sendo assim, o processo de identificação dos sítios de ligação proteína-DNA/RNA pode contribuir para o aprofundamento do conhecimento acerca dessas interações, podendo ser realizado de duas maneiras: utilizando métodos experimentais e por meio de métodos computacionais, como o que será aplicado neste trabalho, que é o aprendizado de máquina (*machine learning*) e *ensemble learning*(SI; ZHAO; WU, 2015; CORSI et al., 2020).

O uso de métodos computacionais está em ascensão nesta área devido ao alto custo e ao tempo de execução dos métodos experimentais, juntamente com o grande número de proteínas para as quais se deseja identificar os sítios de ligação (SI; ZHAO; WU, 2015). O *machine learning* e o *ensemble learning* surgem como ferramentas que possibilitam uma identificação mais ágil e menos custosa, com uma taxa de acertos crescente ao longo dos anos, conforme os estudos na área avançam.

Os métodos para predição de sítios de ligação proteína-DNA/RNA utilizando *machine learning* e *ensemble learning* podem se basear nas informações advindas da sequência de aminoácidos da proteína e da estrutura da proteína e suas características físicas (SI; ZHAO; WU, 2015; XIONG et al., 2018).

1.2 Objetivos

O objetivo deste trabalho é desenvolver um classificador capaz de identificar regiões das proteínas alvo, como exosítios, que modulam a ligação proteína-DNA/RNA, por meio da implementação de algoritmos de aprendizado de máquina e *ensemble learning* utilizando os dados de estruturas de complexos de proteínas e ácidos nucleicos presentes na base *Protein-DNA Docking Benchmark* e os descritores físico-químicos presentes no banco de dados STING_RDB.

1.3 Justificativa

A predição dos sítios de ligação entre proteínas e ácidos nucleicos por meio de técnicas computacionais tem sido amplamente explorada e referenciada em diversos trabalhos, como os de Si, Zhao e Wu (2015), Xiong et al. (2018), no qual os autores abordam algoritmos de *machine learning* e *ensemble learning* que já foram utilizados e apresentaram bons resultados para a tarefa de predição. Em ambos os trabalhos os autores destacam a necessidade de desenvolver métodos computacionais que possam

reduzir os custos e o tempo demandado por métodos experimentais, indicando o uso de técnicas de aprendizado de máquina para a realização da predição dos sítios de ligação.

Métodos computacionais, como o *machine learning* e algoritmos de *ensemble learning*, também são aplicados em trabalhos como o de Wu et al. (2009), Sukumar et al. (2016), Corsi et al. (2020), em que algoritmos de aprendizado de máquinas, juntamente com informações acerca da sequência de aminoácidos, são utilizadas para a predição dos sítios de ligação proteína-DNA, apresentando bom desempenho para a predição dos sítios de ligação proteína-DNA e resultados promissores para sítios de ligação proteína-RNA.

Estes trabalhos demonstram a necessidade de uso e competência dos algoritmos de aprendizado de máquina para a realização da tarefa de predição dos sítios de ligação proteína e ácidos nucleicos.

1.4 Motivação

A identificação de exosítios como moduladores de ligações RNA/DNA pode proporcionar uma maior compreensão do funcionamento celular. Entender as interações proteína-DNA e proteína-RNA é fundamental para se realizar avanços na área de desenvolvimento de fármacos e vacinas (CUI et al., 2022). Embora métodos experimentais possam ser empregados para esta finalidade, eles apresentam altos custos e alta demanda de tempo (SI; ZHAO; WU, 2015; CORSI et al., 2020). Portanto, métodos computacionais, como o aprendizado de máquina e *ensemble learning*, têm sido adotados e possuem resultados promissores (SI; ZHAO; WU, 2015; SUKUMAR et al., 2016).

Capítulo 2

Revisão Bibliográfica

A seção a seguir faz referência aos conceitos teóricos essenciais para a execução deste trabalho.

2.1 Fundamentação Teórica

2.1.1 Proteínas

As proteínas são moléculas que se diferenciam em estrutura, função, sequências de aminoácidos que as compõem, entre outros aspectos. As proteínas são formadas a partir da combinação de 22 aminoácidos e podem apresentar quatro estruturas: primária, secundária, terciária e quaternária (SALIM, 2015).

A estrutura primária é somente a sequência linear de aminoácidos que compõem a proteína. A estrutura secundária é composta por padrões estruturais como as alfa-hélices e folhas beta. A estrutura terciária é a representação tridimensional (3D) da proteína. E, por fim, tem-se a estrutura quaternária, que nem todas as proteínas apresentam, a qual é a junção de várias proteínas (SALIM, 2015).

Dentre as proteínas, podemos destacar o grupo das enzimas. As enzimas participam da catálise de reações, na qual um substrato se liga à enzima para gerar uma reação

enzimática que resulta na transformação do substrato em um novo produto (SALIM, 2015). Essas interações ocorrem através dos sítios de ligação presentes nas enzimas, entre eles podemos destacar os sítios ativos (SALIM, 2015) e os exosítios (BOCK; PANIZZI; VERHAMME, 2007).

Sítios Ativos

Os sítios ativos são regiões da superfície das proteínas que se ligam ao substrato, permitindo que as reações enzimáticas para transformação do mesmo ocorram (SALIM, 2015). Os sítios ativos são constituídos por um grupo de aminoácidos que facilitam as interações necessárias para a ligação da proteína com o substrato. Além disso, os sítios contêm os resíduos de aminoácidos catalíticos, os quais participam das quebras e ligações químicas necessárias para a geração do produto (SALIM, 2015).

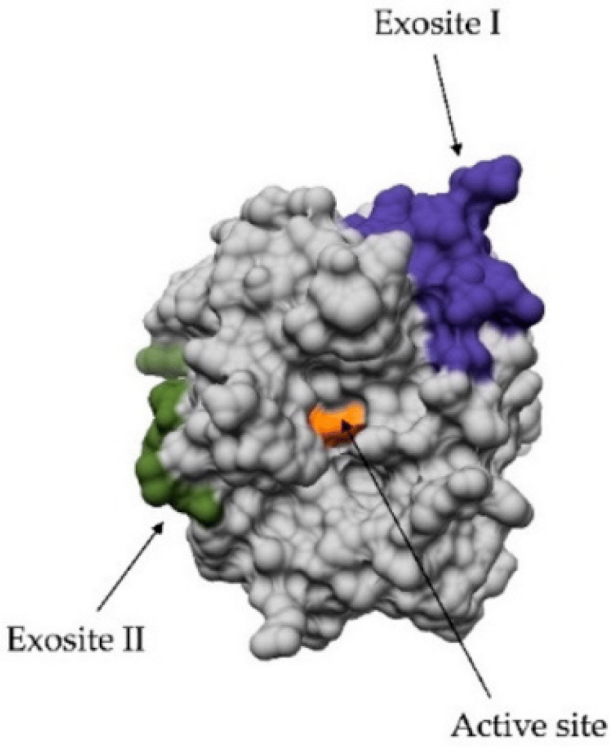
Outros sítios, como os exosítios, que são o foco deste trabalho, também desempenham papéis muito importantes nas ligações e interações realizadas pelas proteínas (BOCK; PANIZZI; VERHAMME, 2007; OBAHA; NOVINEC, 2023).

Exosítios

Os exosítios são regiões na superfície das proteínas que funcionam como sítios de ligação secundários (YEGNESWARAN et al., 2007; BOCK; PANIZZI; VERHAMME, 2007). Esses sítios se localizam afastados do sítio ativo, mas ainda são muito importantes para a ligação com diferentes tipos de ligantes (BOCK; PANIZZI; VERHAMME, 2007; OBAHA; NOVINEC, 2023). Desse modo, juntamente com os sítios ativos, os exosítios também têm um papel importante na definição da afinidade de ligação entre proteínas e substratos (OBAHA; NOVINEC, 2023; BOCK; PANIZZI; VERHAMME, 2007). Devido à sua relevância para as ligações, os exosítios se tornam alvo de trabalhos que buscam possibilitar a criação de novos fármacos (OBAHA; NOVINEC, 2023).

A representação de uma proteína, seu sítio ativo e exosítios pode ser visualizada na Figura 2.1.

Figura 2.1: Representação da proteína trombina com os exosítios e o sítio ativo



Fonte: (OBAHA; NOVINEC, 2023)

Interações proteína-DNA/RNA

As interações proteína-DNA/RNA desempenham diversos papéis na manutenção da vida, sendo o centro de vários processos biológicos (CUI et al., 2022; CORSI et al., 2020). Estando relacionadas a diversas doenças tornam-se alvos terapêuticos (CORSI et al., 2020) .

Ao aprofundar o conhecimento sobre essas interações, espera-se um maior entendimento dos mecanismos celulares envolvidos, o que pode levar à criação de novas drogas e vacinas (CUI et al., 2022).

Tendo isso em vista, trabalhos buscam identificar os sítios de ligação entre pro-

teínas e ácidos nucleicos utilizando informações acerca da sequência e estrutura das proteínas a fim de possibilitar este aprofundamento sobre o assunto e possibilitar o desenvolviemnto de fármacos a partir destas informações (CORSI et al., 2020; SUKUMAR et al., 2016; WU et al., 2009).

2.1.2 Aprendizado de Máquina

Aprendizado de máquina ou *machine learning* é uma das áreas de estudo do campo de inteligência artificial (IA) que abrange algoritmos capazes de aprender com base em conjuntos de treinamento e a partir disso realizar tarefas como predições, escolhas, classificações, entre outras (SHINDE; SHAH, 2018). Os algoritmos são divididos em três tipos de acordo com a metodologia de treinamento, sendo eles: supervisionados, não supervisionados e semi-supervisionados (SALIM, 2015).

Os algoritmos supervisionados têm como característica o uso de conjuntos de treinamento com dados previamente rotulados, ou seja, já classificados corretamente. Já no aprendizado não supervisionado, as entradas não são rotuladas, sendo o resultado desconhecido em primeiro momento. E os métodos semi-supervisionados, por sua vez, são uma abordagem intermediária, em que existem entradas com e sem rótulos (SALIM, 2015).

A escolha do método de aprendizado, assim como do algoritmo a ser utilizado depende das características dos dados de entrada, tanto para treinamento quanto validação, e da natureza do problema a ser resolvido (CUI et al., 2022; SI; ZHAO; WU, 2015; SALIM, 2015).

Aprendizado Supervisionado

Neste trabalho, optou-se por utilizar algoritmos de aprendizado supervisionado, sendo alguns exemplos de algoritmos: *Linear Classifiers*, *Logistic Regression*, *Perceptron*, SVM, entre outros (OSISANWO et al., 2017). Dentre os algoritmos de

aprendizado supervisionado, podemos destacar o algoritmo de *decision tree* ou árvore de decisão utilizado em tarefas de *data mining* e *machine learning* (RANA et al., 2014), que utiliza um modelo que se assemelha à estrutura de uma árvore, em que a classificação começa na raiz e segue até as folhas, sendo os nós e ramos percorridos referenciando as *features* e os valores que podem ter (OSISANWO et al., 2017; SI; ZHAO; WU, 2015). Para cada entrada, afim de se realizar a classificação as *features* são avaliadas nos nós e dependendo do valores são percorridos os ramos até chegarem a uma classificação representada por uma das folhas da árvore (SHARMA; KUMAR et al., 2016). Exemplos desse tipo de algoritmo incluem CHID, CART, ID3 e C4.5 (RANA et al., 2014).

Neste trabalho, serão utilizados algoritmos de *ensemble learning* que combinam múltiplas árvores de decisão a fim de realizar a predição dos exosítios que modulam as ligações proteína-DNA/RNA.

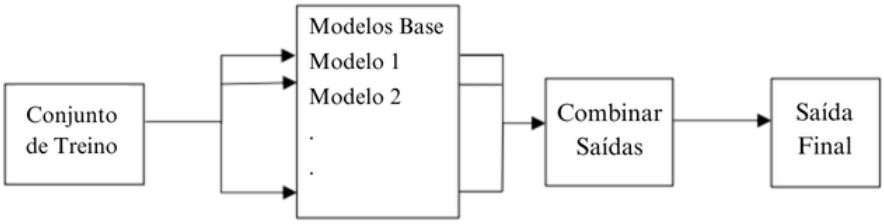
2.1.3 *Ensemble Learning*

Ensemble learning é um termo que faz referência a algoritmos que são a combinação de métodos de *machine learning*. Seu conceito principal vem do conhecimento coletivo em que se espera que os erros de um classificador sejam compensados pelos outros que fazem parte do algoritmo (SAGI; ROKACH, 2018).

No desenvolvimento dos modelos de aprendizado de máquina que compõem o método de *ensemble*, podem ser aplicadas dois tipos de estruturas: a primeira é a de forma independente, em que a construção de um modelo não afeta os outros, ou, então, de forma dependente em que as saídas dos classificadores irão afetar na implementação dos demais (SAGI; ROKACH, 2018).

Ao se utilizar *ensemble learning*, espera-se que se possa ter um desempenho maior do que os algoritmos de aprendizado de máquina podem alcançar separadamente, a arquitetura básica de um modelo de *ensemble learning* está ilustrada na Figura 2.2.

Figura 2.2: Arquitetura básica de métodos de *ensemble learning*



Fonte: Adaptado de (MAHAJAN et al., 2023)

2.1.4 Algoritmos

Neste trabalho, foram utilizados algoritmos de *ensemble learning* que combinam métodos de aprendizado de máquina supervisionado para a construção do modelo. Como exemplos, podemos citar os explorados neste trabalho: *Random Forest*, *LightGBM* e *XGBoost*.

Random Forest

O algoritmo *Random Forest* é um algoritmo de *ensemble learning* supervisionado que tem como base o uso de múltiplas árvores de decisão para realizar a classificação (BOULESTEIX et al., 2012; BELGIU; DRĂGUŢ, 2016). O modelo é capaz de detectar associações não lineares e demonstra bom desempenho lidando com conjuntos de dados que têm poucas observações, mas que apresentam diversas variáveis (BOULESTEIX et al., 2012).

O algoritmo já foi usado em outros trabalhos para previsão de sítios de proteínas (WU et al., 2009), além de que para sua escolha também foram realizados estudos sobre algoritmos que teriam bom desempenho com os dados dos descritores.

LightGBM

LightGBM é um algoritmo de *ensemble learning* recente, que tem apresentado ótimos resultados, além de um baixo custo computacional. É um algoritmo de *Gradient*

Boosted Decision Tree (GBDT), que usa algoritmos de árvores de decisão como base (KE et al., 2017; MIENYE; SUN, 2022). Ele faz uso de um método baseado em histogramas para obter um melhor tempo de treinamento e uma redução no consumo de memória (MACHADO; KARRAY; SOUSA, 2019). O algoritmo *LightGBM* foi selecionado após estudos sobre algoritmos que lidariam bem com o banco de dados.

XGBoost

O *XGBoost* é um algoritmo de *ensemble learning* supervisionado baseado em árvores de decisão, que implementa a estrutura de *gradient boosting*, se sobressaindo pela sua rapidez, seu potencial de lidar com ausência de dados e não necessitar de uma implementação profunda de *feature selection*, apresentando um bom desempenho para tarefas de classificação (MIENYE; SUN, 2022). O *XGBoost* foi escolhido após estudos sobre algoritmos que funcionariam bem com o banco de dados de descritores físico-químicos das proteínas.

2.1.5 STING_RDB

O STING_RDB é um banco de dados mantido pela Embrapa, sendo uma solução criada a partir do STING_DB, que é uma das maiores bases de dados de descritores estruturais de proteínas. Entretanto, o STING_DB apresenta dificuldades para grandes análises de dados, devido aos dados estarem dispostos em vários arquivos com formatações diversas (SALIM, 2015). Sendo assim, foi desenvolvido o STING_RDB para facilitar este processo, reunindo os dados em um banco relacional (SALIM, 2015; OLIVEIRA et al., 2007).

Estando conectado com bases como PDB, UniProt, entre outras, o STING_RDB recebe atualizações semanais que adicionam novas estruturas. Sendo assim, os descritores destas estruturas são calculados e adicionados à base (SALIM, 2015; OLIVEIRA et al., 2007).

2.1.6 *Protein-DNA Docking Benchmark*

O *Protein-DNA Docking Benchmark* é uma base de dados que contém 47 estruturas de complexos entre proteínas e ácidos nucleicos (DIJK; BONVIN, 2008), abrangendo a maioria dos grupos de proteínas que se ligam a ácidos nucleicos. O *benchmark* é dividido em três categorias: *easy*, *intermediate* e *difficult*, com 13, 22 e 12 estruturas, respectivamente (DIJK; BONVIN, 2008; RODRÍGUEZ-LUMBRERAS et al., 2022). Por não apresentar redundâncias e abranger diversos grupos de proteínas, essa base torna-se uma ferramenta importante para o desenvolvimento de métodos computacionais envolvendo os sítios de ligação entre proteínas e ácidos nucleicos (DIJK; BONVIN, 2008).

2.1.7 *Feature Selection*

Feature selection ou seleção de características é uma etapa de pré processamento de dados em que é realizada a a seleção das *features* relevantes para o processo de classificação eliminando características que podem afetar e diminuir o desempenho do modelo (TANG; ALELYANI; LIU, 2014; FURLANETTO; GOMES; BREVE, 2023). Com a diminuição na quantidade de características há também uma redução no custo computacional e melhora na qualidade das predições (MIAO; NIU, 2016). Esse processo pode ser feito atraves de três categorias de métodos: *filter*, *wrapped* e *embedded*.

Nos métodos *filter* a seleção de características é realizada sem levar em conta o algoritmo de aprendizado, visando somente as especifidade dos dados. O método *filter* apresenta um custo computacional baixo, além de ter uma implementação simples, entretanto, ignora a natureza dos algoritmos de *machine learning* (FURLANETTO; GOMES; BREVE, 2023; TANG; ALELYANI; LIU, 2014). Como exemplos de técnicas *filter*, estão o uso do teste qui-quadrado e baseado em correlação.

- Teste qui-quadrado: O teste qui-quadrado foi aplicado neste trabalho e é utilizado para calcular a independência entre duas *features*. Em uma abordagem de

seleção de características utilizando métodos *filter*, o qui-quadrado é empregado para avaliar o valor de independência entre uma característica e os alvos, que são as classes. O objetivo é selecionar as características mais dependentes, que são consideradas mais relevantes para o processo de classificação (THASEEN; KUMAR, 2017; THASEEN; KUMAR; AHMAD, 2019).

- *Feature selection* baseado em correlação: O método baseado em correlação realiza o cáculo do valor de correlação entre as *features* e a classe, como também entre as próprias características. *Features* com baixo valor de correlação com as classes são eliminadas e consideradas irrelevantes, além disso, *features* que têm um valor de correlação muito alto entre si são redundantes não trazendo informação adicional para a tarefa de classificação e também são excluídas. Desta forma, as *features* importantes para o processo de classificação são selecionadas (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2013).

Métodos *wrapper*, diferentemente de *métodos filter*, utilizam o algoritmo de aprendizado no processo de seleção, que realizará a função de avaliar os possiveis conjuntos de *features*, selecionando o que obtiver maior pontuação. O problema desse método está associado ao grande custo computacional para grandes bases de dados (FURLANETTO; GOMES; BREVE, 2023). Exemplos de métodos *wrapper* são *Recursive Feature Elimination* e *Sequential Feature Selection* (GRANITTO et al., 2006; AGGRAWAL; PAL, 2020).

- *Recursive Feature Elimination* (RFE): O RFE é um método *wrapper* aplicado neste trabalho que utiliza um processo recursivo em que, a cada iteração, a importância de cada *feature* é calculada por meio da execução de um modelo. Em cada iteração, a *feature* ou um grupo de *features* de tamanho estipulado com o menor valor de *feature importance* para classificação são eliminados, de forma que, nas próximas iterações, eles não estejam presentes. O processo continua até que se atinja um número determinado de *features* (GRANITTO et al., 2006).

- *Sequential Feature Selection* (SFS): Neste método, o conjunto de *features* selecionadas começa vazio. Na primeira iteração, todas as *features* são testadas separadamente em um modelo de classificação até encontrar aquela que gerou o melhor desempenho do classificador, essa *feature* é então adicionada ao conjunto, nas próximas iterações as *features* restantes fora do conjunto são testadas uma a uma junto com o conjunto de características selecionadas e a que apresentar o maior desempenho é adicionada ao conjunto. O processo se repete até atingir um número de caracteristicas determinado (AGGRAWAL; PAL, 2020).

Os métodos *embedded* são um meio termo entre os métodos *filter* e *wrapper*. No *embedded*, a seleção de *features* é realizada ao mesmo tempo que o treino do modelo, dessa forma levando a um custo computacional menor do que o *wrapped* (TANG; ALELYANI; LIU, 2014), exemplos de métodos embedded são *feature selection Tree-Based* e baseada em *perceptron* (ALSAHAF et al., 2022; BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2013).

- *Tree-Based*: Um dos métodos de *feature selection* abordados neste trabalho foi o *tree-based*, uma abordagem *embedded* em que um modelo baseado em árvores de decisão é treinado. A partir das métricas de *feature importance* geradas pelo modelo, as *features* que tiveram maior importância para a classificação são selecionadas (ALSAHAF et al., 2022).

- *Feature selection* baseada em *perceptron*: O método *embedded* baseado em *perceptron* constitui no treino do modelo de rede neural com a intenção de avaliar os pesos das interconexões formadas, que são utilizados para selecionar as *features* mais relevantes para o problema de classificação (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2013).

2.1.8 Métricas

Para avaliaçao de modelos de aprendizado de máquina e *ensemble learning* são utilizadas métricas afim de avaliar o desempenho de classificadores (ERICKSON; KITAMURA, 2021). São utilizadas para isso métricas como acurácia, *F1-Score*, precisão e *recall* (ERICKSON; KITAMURA, 2021). As fórmulas utilizadas para o cálculo destas métricas são apresentadas a seguir, onde VP representa os verdadeiros positivos, VN os verdadeiros negativos, FP os falsos positivos e FN os falsos negativos.

- Acurácia:

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{2.1}$$

- Precisão:

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.2}$$

- *Recall*:

$$\textit{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.3}$$

- *F1-Score*:

$$\textit{F1-Score} = 2 \times \frac{\textit{Precisão} \times \textit{Recall}}{\textit{Precisão} + \textit{Recall}} \tag{2.4}$$

Além destas, outras métricas são utilizadas, como por exemplo a curva ROC, que representa o balanço entre a taxa de verdadeiros positivos e falsos positivos, a AUC, que é a área sob a curva ROC, a AUCPRC, área sob a curva precisão-*recall* (ERICKSON; KITAMURA, 2021) e MCC , *Matthews correlation coefficient*, que indica a correlação entre a classe real e a que foi predita pelo classificador (CHICCO; TÖTSCH; JURMAN, 2021).

2.2 Trabalhos Correlatos e Estado da Arte

Em trabalhos como o desenvolvido por Sukumar et al. (2016), é implementado o modelo DBSI, uma aplicação que utiliza o algoritmo SVM (*Support Vector Machine*), uma técnica de aprendizado de máquina supervisionado, para identificação de sítios de ligação proteína-DNA através das características da sequência de aminoácidos e da estrutura das proteínas. O modelo apresentou uma acurácia de 0,83 e *F1-Score* de 0,59 para a predição de sítios de ligação proteína-DNA, além de resultados promissores na predição de sítios de ligação proteína-RNA com uma acurácia de 0,77 e *F1-Score* de 0,59, mostrando que o uso de métodos computacionais é imprescindível para reduzir o tempo e os custos associados à identificação desses sítios.

No trabalho de Corsi et al. (2020) também é proposto o uso de métodos computacionais para a predição dos sítios de ligação proteína-DNA. Nele é descrita a ferramenta $\text{JET}^2_{\text{DNA}}$, a qual foi desenvolvida utilizando o *support-core-rim model* para a tarefa de predição. A ferramenta utiliza três descritores para realizar as predições. Os descritores usados apresentam a significância biológica e física necessárias para a aplicação. O $\text{JET}^2_{\text{DNA}}$ apresentou uma sensibilidade e *F1-Score* aproximadamente 20% e 5% maior respectivamente que outros classificadores utilizados para a predição de sítios de ligação proteína-DNA e ao ser aplicada para a predição de sítios de ligação proteína-RNA, a ferramenta não teve perda de desempenho, mostrando que os sítios de ligação proteína com DNA e RNA apresentam similaridades.

No trabalho desenvolvido por Wu et al. (2009) é implementado o algoritmo *Random Forest* para predição de sítios de ligação proteína-DNA devido à sua agilidade e robustez. A abordagem utilizada combina informações sobre a sequência de aminoácidos com uma *feature* híbrida que contém informações evolutivas da sequência, dados da estrutura secundária e informações do vetor binário ortogonal, sem recorrer a informações da estrutura 3D da proteína para realização das predições. Os autores realizam também a comparação com o algoritmo SVM para a predição dos sítios de

ligação. Os resultados mostram que ambos os métodos apresentam bom desempenho para a tarefa, tendo o *Random Forest* o valor de AUC de 0,913 e o SVM de 0,921, entretanto, o *Random Forest* é muito mais rápido e robusto, sendo indicado pelo estudo como um método mais efetivo para predição dos sítios de ligação proteína-DNA.

Em trabalhos recentes da literatura como o de Zeng et al. (2020), técnicas computacionais têm sido implementadas para predição de sítios de ligação de proteínas. Neste artigo, os autores abordam o desenvolvimento do modelo DeepPPISP, que utiliza *deep learning* juntamente com a combinação de *features* locais e *features* globais da sequência de aminoácidos da proteína para realizar a predição de sítios de ligação proteína-proteína. A utilização de ambas *features* locais e globais diferencia o DeepPPISP de outros modelos que realizam a predição baseando-se somente nas características locais. O resultado obtido dessa combinação é um modelo que obteve métricas como *F1-Score* e MCC superiores aos modelos concorrentes PSIVER, SPPIDER, SPRINGS, ISIS e RF_PPI, mas que sofre com alto tempo de execução, não sendo ideal para proteínas com longas sequências de aminoácidos.

A identificação dos sítios de ligação de proteínas também foi abordada por Li, Golding e Ilie (2021). Neste trabalho, é relatada a implementação do DELPHI, uma ferramenta de predição de sítios de ligação proteína-proteína que usa *deep learning*. O modelo realiza a tarefa de predição utilizando apenas informações advindas da sequência de aminoácidos, por meio de uma abordagem na qual uma rede neural convolucional (CNN) e uma rede neural recorrente (RNN) são combinadas. Além disso, o modelo utiliza três *features* que não foram abordadas por outros trabalhos da área no momento da publicação. O DELPHI apresentou valores de métricas como sensibilidade, especificidade, acurácia, *F1-Score*, MCC, AUROC e AUPRC maiores quando comparados com outras 9 ferramentas do estado da arte utilizadas para predição dos sítios, validando o uso de *ensemble*, que foi capaz de aumentar o desempenho do modelo.

Utilizar técnicas de aprendizado de máquina e *ensemble learning* também foi uma

abordagem adotada por Johansson-Åkhe, Mirabello e Wallner (2019). Neste trabalho, os autores implementaram uma *pipeline* nomeada InterPep, que utiliza os algoritmos de *Random Forest* e *hierarchical clustering* para a predição de sítios de interação proteína-peptídeos. A InterPep utiliza estruturas *template* e informações obtidas da sequência de aminoácidos dos peptídeos para realizar a predição de locais com maior chance de que a ligação proteína-peptídeo ocorram. A pipeline alcançou uma precisão de 80% juntamente com um *recall* de 20% demonstrando ser um bom ponto de partida para experimentos que buscam entender as ligações entre porteínas e peptídeos.

Capítulo 3

Desenvolvimento e Plano de Trabalho

Nesta seção, serão abordadas as atividades realizadas para o desenvolvimento deste trabalho.

3.1 Desenvolvimento do Trabalho

Inicialmente, para realização deste trabalho, estudos envolvendo proteínas, exosítios e a interação proteína e ácidos nucleicos foram realizados com o objetivo de compreender melhor os conceitos necessários para a execução do trabalho. Além disso, foram feitos estudos sobre métodos de aprendizado de máquina e *ensemble learning* que serão utilizados para identificar exosítios que modulam as ligações proteína-DNA e proteína-RNA, como também sobre métodos de pré-processamento de dados, como *undersampling* e *feature selection*. A criação do banco, o tratamento dos dados, a seleção e a implementação dos métodos estão descritos nas subseções a seguir.

3.1.1 Banco de Dados

Para a implementação dos modelos, foi necessário criar uma banco de dados para o treinamento e testes, contendo resíduos das superfícies de proteínas que fazem parte das interfaces de ligação proteínas e ácidos nucleicos, assim como resíduos que não

pertencem a essas interfaces. Esses dois grupos compõem as classes do problema de classificação.

Para este trabalho, foi utilizada a base de dados *Protein-DNA Docking Benchmark*, que contém 47 estruturas de complexos proteína e ácidos nucleicos (RODRÍGUEZ-LUMBRERAS et al., 2022), juntamente com a informação vindas dos descritores físico-químicos do STING_RDB. Inicialmente, foi utilizado um *script* para separar as proteínas das cadeias de ácidos nucleicos nos complexos do *Protein-DNA Docking Benchmark*. Em seguida, os mesmos descritores presentes no STING_RDB foram calculados para as estas proteínas aplicando os mesmos *softwares*, gerando um banco de dados denominada STING_Prot_DNA_Benchmark_bound, que serviu de base para a criação do banco ProtDNA_STINGRDB2_Descritores utilizada neste trabalho para o treino e teste dos modelos.

A partir do banco STING_Prot_DNA_Benchmark_bound, foram selecionados os resíduos localizados na superfície das proteínas e 284 descritores relevantes para a tarefa de classificação, vindos das tabelas *Residue, Cross_Link_Order, Cross_Presence_Order, Curvature, Density_Sponge, Electrostatic_Potential, Energy_Density, Graph_Descriptor, Hydrophobicity, Solvation, Unused_Contacts, Weighted_Contact_Number, Residue_Contacts e Air*, que foram então usados para compor uma tabela presente no banco ProtDNA_STINGRDB2_Descritores. Utilizou-se o descritor *acc_isol_surfv* como critério para identificar os resíduos de superfície, selecionando os resíduos com valores de *acc_isol_surfv* maior que 0. Para definir as classes, ou seja, determinar se os resíduos fazem ou não parte dos sítios de ligação proteínas e ácidos nucleicos, foi realizada uma comparação entre as informações dos resíduos presentes no banco ProtDNA_STINGRDB2_Descritores, derivado da *Protein-DNA Docking Benchmark* após remoção dos ácidos nucleicos, e os dados do STING_RDB.

A classificação dos resíduos do banco ProtDNA_STINGRDB2_Descritores foi

feita utilizando o descritor *acc_complex_surfv*, que é um descritor de acessibilidade do resíduo, que é menor quando ele esta em contato com o DNA ou RNA. O valor de cada resíduo foi comparado ao valor correspondente no STING_RDB, onde os descritores foram calculados com a presença dos ácidos nucleicos. Portanto, ao comparar os valores entre o STING_RDB e o banco criado para este trabalho, se os valores forem iguais, isso indica que o resíduo não faz parte da interface de ligação, sendo classificado como 0. No entanto, se o valor presente na base ProtDNA_STINGRDB2_Descritores for maior do que o do STING_RDB, isso significa que o resíduo faz parte da interface, e que sua acessibilidade aumentou em relação ao STING_RDB quando a cadeia de ácidos nucleicos foi removida, indicando que o resíduo interage com as cadeias, sendo classificado como 1. O banco criado para este trabalho contém 13.830 entradas, sendo 3.227 referentes à classe 1 e 10.603 à classe 0.

3.1.2 Pré-Processamento

Em problemas de classificação, como o abordado neste estudo, é comum encontrar classes desbalanceadas, o que pode prejudicar o desempenho dos classificadores (LEE; SEO, 2022). Para mitigar esse problema, foi aplicado o método de *undersampling*, que consiste em reduzir a quantidade de amostras da classe majoritária. Esse processo visa equilibrar as classes e, assim, melhorar o desempenho dos modelos (SHELKE; DESHMUKH; SHANDILYA, 2017). Esse método foi utilizado para igualar o tamanho da classe 0 de resíduos que não fazem parte dos sítios de ligação proteína-DNA e proteína-RNA com o número de entradas da classe 1 de resíduos que participam da ligação proteína e ácidos nucleicos. Sendo assim, foi obtido um *dataset* com 6.454 entradas, sendo 3.227 da classe 1 e 3.227 da classe 0.

Em seguida o processo de *feature selection* foi realizado com o objetivo de identificar as *features* relevantes para a classificação, melhorando o desempenho dos modelos

e reduzindo o custo computacional.

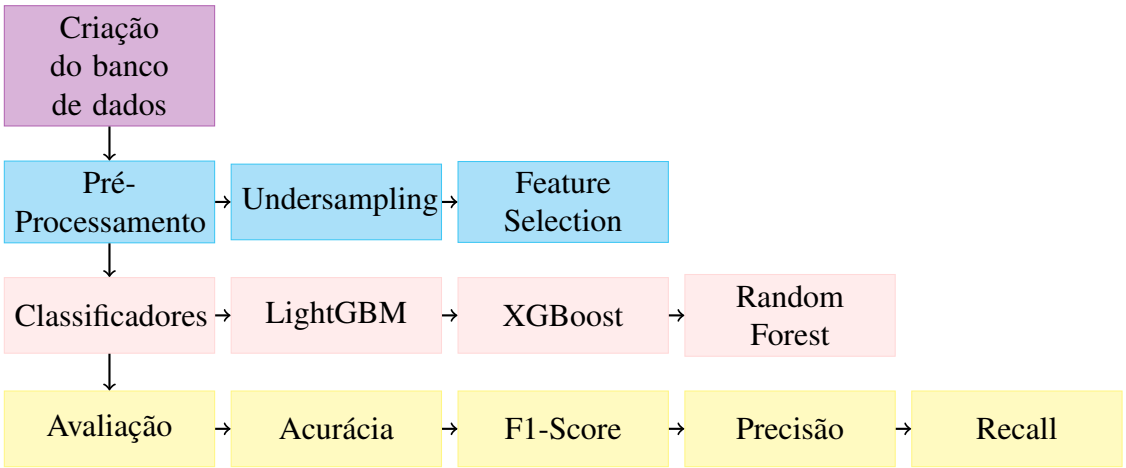
3.1.3 Implementação dos Métodos

A implementação dos métodos se deu primeiramente pela identificação de algoritmos que já lidaram bem com os dados de descritores físico-químicos do STING_RDB em trabalhos anteriores e que já foram utilizados previamente para a predição de sítios de ligação entre proteínas e ácidos nucleicos (WU et al., 2009). Sendo assim, foram escolhidos os algoritmos *Random Forest*, *LightGBM* e *XGBoost* para serem explorados neste trabalho.

Os algoritmos fazem uso de técnicas de *ensemble learning*, baseando-se no uso de múltiplas árvores de decisão para realizar a predição. Os modelos foram implementados utilizando a linguagem de programação *Python* devido às bibliotecas de *machine learning* como *scikit-learn*, *lightgbm* e *xgboost* disponíveis para uso.

Modelos de classificação a partir desses algoritmos foram gerados, treinados e testados com seus resultados sendo avaliados a fim de identificar aquele que obteve o melhor desempenho para a tarefa de predição dos exosítios. As etapas para implementação dos modelos pode ser visualizada na Figura 3.1.

Figura 3.1: Diagrama da implementação dos modelos



Fonte: Elaborado pelo autor.

Capítulo 4

Testes e Resultados

Nesta seção, são abordados os testes realizados e os resultados obtidos pelos métodos de *ensemble learning*.

4.1 Realização dos testes

Os testes foram realizados utilizando o banco de dados criado para este trabalho, o ProtDNA_STINGRDB2_Descritores. Em cada teste, foi aplicado o *undersampling* para balanceamento das classes, assim como uma técnica de seleção de características que foram: um método *embedded*, o *tree-based*, além de um método *wrapped* o *Recursive Feature Elimination*(RFE) e uma técnica *filter*, o teste qui-quadrado. Em seguida, os modelos foram treinados e testados com as *features* selecionadas por cada método de *feature selection*, seguindo a estratégia de *10-fold cross validation*, está foi escolhida por ser uma das mais utilizadas por proporcionar um balanço entre custo computacional e a avaliação do desempenho do modelo (PACHOULY et al., 2022), consistindo na divisão aleatória do banco em dez conjuntos. Esses conjuntos foram utilizados em 10 iterações de treino e teste, nas quais, em cada iteração, nove desses conjuntos gerados foram usados para treino e um para teste, mudando a cada iteração o conjunto utilizado para teste, assegurando que a cada etapa de treino e teste um con-

junto diferente seja usado para teste. Deste modo, para cada iteração, foram obtidas as métricas de acurácia, *F1-Score*, precisão e *recall weighted* de cada modelo, que tiveram suas médias e desvio padrão calculados, juntamente com as matrizes de confusão de cada um dos modelos para cada técnica de *feature selection*.

Os parâmetros de cada algoritmo de classificação foram:

- 1 - *LightGBM*: 100 árvores de decisão, *Gradient Boosting Decision Tree* como algoritmo de *boosting* e taxa de aprendizado de 0,1.
- 2 - *XGBoost*: 100 árvores de decisão, *Gradient Boosting Tree* como o *booster* e taxa de aprendizado de 0,3.
- 3 - *Random Forest*: 100 árvores de decisão.

4.2 Resultados

Nas seções a seguir, serão apresentados os resultados de cada teste conduzido, considerando os diferentes métodos de seleção de características e classificadores.

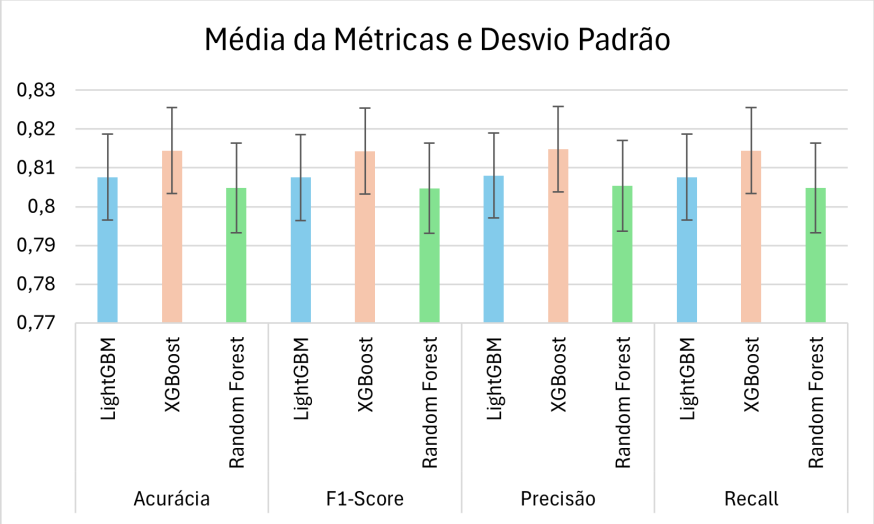
4.2.1 *Feature Selection Tree-Based*

Os resultados obtidos dos modelos utilizando o método *tree-based* para seleção das características usadas para treino e teste podem ser visualizados na Tabela 4.1, bem como na Figura 4.1 que ilustra os valores das métricas de cada modelo permitindo melhor comparação entre os classificadores, as matrizes de confusão dos modelos podem ser visualizadas nas tabelas: Tabela 4.2, Tabela 4.3 e Tabela 4.4.

Tabela 4.1: Resultados dos modelos de aprendizado de máquina com *feature selection tree-based*

Modelo	Métrica	Média	Desvio Padrão
LightGBM	Acurácia	0,8076	0,0111
	F1-Score	0,8075	0,0111
	Precisão	0,8080	0,0110
	Recall	0,8076	0,0111
XGBoost	**Acurácia**	**0,8144**	**0,0111**
	F1-Score	**0,8143**	**0,0111**
	Precisão	**0,8148**	**0,0110**
	Recall	**0,8144**	**0,0111**
Random Forest	Acurácia	0,8048	0,0116
	F1-Score	0,8047	0,0116
	Precisão	0,8053	0,0117
	Recall	0,8048	0,0116

Figura 4.1: Média das métricas e desvio padrão utilizando o método *Tree-Based*



Fonte: Elaborado pelo próprio autor.

Tabela 4.2: Matriz de Confusão - *LightGBM* - *Tree-Based*

	Previsto Positivo	Previsto Negativo
Real Positivo	2563	664
Real Negativo	578	2648

Tabela 4.3: Matriz de Confusão - *XGBoost* - *Tree-Based*

	Previsto Positivo	Previsto Negativo
Real Positivo	2580	647
Real Negativo	551	2676

Tabela 4.4: Matriz de Confusão - *Random Forest* - *Tree-Based*

	Previsto Positivo	Previsto Negativo
Real Positivo	2556	671
Real Negativo	589	2638

Neste teste foi realizada a seleção de *features* baseada em árvores com a utilização do algoritmo *Random Forest* com 100 árvores que foi aplicado, e as *features* com valor de *feature importance* maior que a média da importância de todas as características foram selecionadas para o processo de classificação, eliminando da base aquelas que não eram relevantes. Em seguida, os modelos foram treinados e testados, por meio da técnica de *10-fold cross validation* e suas métricas foram calculadas. Nota-se que o algoritmo que mais se destacou foi o *XGBoost* ao apresentar os maiores valores médios de acurácia, *F1-Score*, precisão e *recall*, o que pode ser explicado por sua característica de utilizar *Gradient Boosting Tree* como *booster*. Com as segundas maiores médias ficou o algoritmo *LightGBM*, e por último, o *Random Forest* com os menores valores.

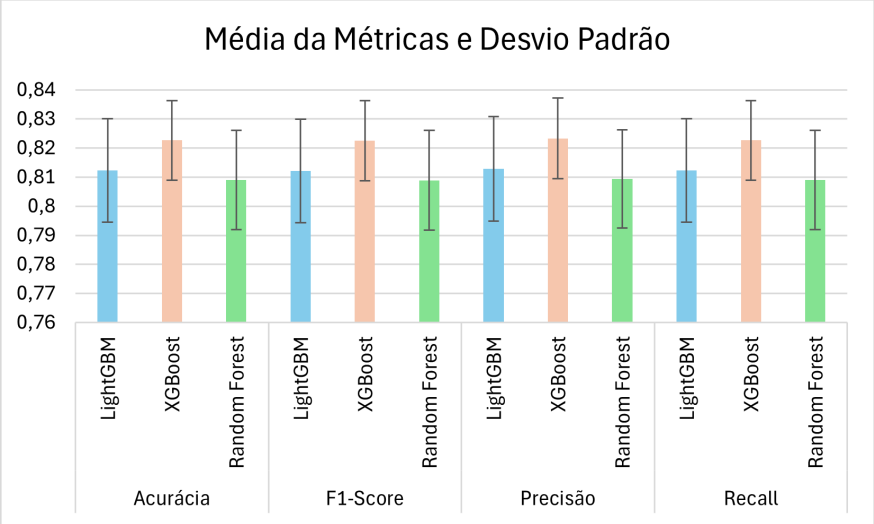
4.2.2 Feature Selection utilizando Recursive Feature Elimination (RFE)

Na Tabela 4.5 podem ser observadas as métricas médias obtidas por cada classificador após a realização de *feature selection* utilizando o RFE, assim como na Figura 4.2 em que é possivel realizar uma melhor comparação do desempenho dos modelos, as matrizes de confusão dos modelos podem ser vistas nas tabelas: Tabela 4.6, Tabela 4.7 e Tabela 4.8. .

Tabela 4.5: Resultados dos modelos de aprendizado de máquina com *feature selection* utilizando RFE

Modelo	Métrica	Média	Desvio Padrão
LightGBM	Acurácia	0.8122	0,0178
	F1-Score	0,8121	0,0178
	Precisão	0,8128	0,0180
	Recall	0,8122	0,0178
XGBoost	**Acurácia**	**0,8226**	**0,0137**
	F1-Score	**0,8225**	**0,0137**
	Precisão	**0,8233**	**0,0139**
	Recall	**0,8226**	**0,0137**
Random Forest	Acurácia	0,8090	0,0170
	F1-Score	0,8089	0,0171
	Precisão	0,8094	0,0169
	Recall	0,8090	0,0170

Figura 4.2: Média das métricas e desvio padrão utilizando o método RFE



Fonte: Elaborado pelo próprio autor.

Tabela 4.6: Matriz de Confusão - *LightGBM* - RFE

	Previsto Positivo	Previsto Negativo
Real Positivo	2580	647
Real Negativo	565	2662

Tabela 4.7: Matriz de Confusão - *XGBoost* - RFE

	Previsto Positivo	Previsto Negativo
Real Positivo	2598	629
Real Negativo	516	2711

Tabela 4.8: Matriz de Confusão - *Random Forest* - RFE

	Previsto Positivo	Previsto Negativo
Real Positivo	2585	642
Real Negativo	591	2636

O teste foi realizado com uso do processo de seleção de *features* utilizando RFE por meio da avaliação da importância das *features* para classificação obtidas apartir do treino de um modelo de *Random Forest* com 10 árvores eliminado por iteração a característica com o menor valor de importância até restar 150 que forma selecionadas para o treino e teste. Os modelos foram treinados e testados seguindo o *10-fold cross validation*, e suas métricas foram calculadas. Neste teste, o modelo *XGBoost* apresentou o melhor desempenho, alcançando as maiores médias de acurácia, *F1-Score*, precisão e *recall*, isso se deve a sua capacidade de utilizar as o *Gradient Boosting Tree* e combinar as saídas das múltiplas árvores de decisão. O *LightGBM* teve o segundo melhor desempenho neste teste, seguido pelo *Random Forest* com as métricas mais baixas dentre os três.

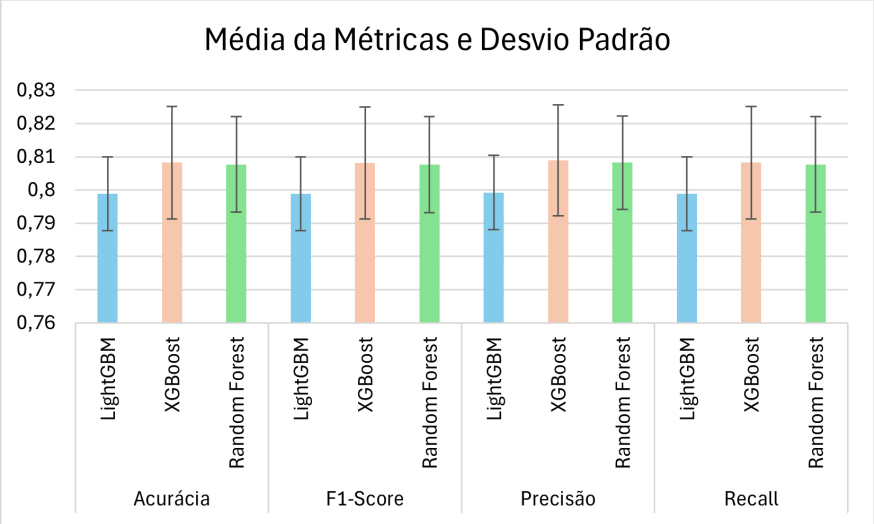
4.2.3 *Feature Selection* utilizando o teste qui-quadrado

Os resultados obtidos utilizando o teste qui-quadrado para seleção das *features* pode ser visualizado na Tabela 4.9, juntamente com a Figura 4.3, em que são representados os resultados de desempenho de cada classificador permitindo uma melhor comparação. Além disso, nas tabelas: Tabela 4.10, Tabela 4.11 e Tabela 4.12, encontram-se as matrizes de confusão dos modelos. .

Tabela 4.9: Resultados dos modelos de aprendizado de máquina com *feature selection* utilizando o teste qui-quadrado

Modelo	Métrica	Média	Desvio Padrão
LightGBM	Acurácia	0,7989	0,0111
	F1-Score	0,7988	0,0111
	Precisão	0,7992	0,0112
	Recall	0,7989	0,0111
XGBoost	**Acurácia**	**0,8082**	**0,0169**
	F1-Score	**0,8081**	**0,0169**
	Precisão	**0,8089**	**0,0167**
	Recall	**0,8082**	**0,0169**
Random Forest	Acurácia	0,8077	0,0144
	F1-Score	0,8076	0,0145
	Precisão	0,8082	0,0141
	Recall	0,8077	0,0144

Figura 4.3: Média das métricas e desvio padrão utilizando o teste qui-quadrado



Fonte: Elaborado pelo próprio autor.

Tabela 4.10: Matriz de Confusão - *LightGBM* - Teste Qui-Quadrado

	Previsto Positivo	Previsto Negativo
Real Positivo	2554	673
Real Negativo	625	2602

Tabela 4.11: Matriz de Confusão - *XGBoost* - Teste Qui-Quadrado

	Previsto Positivo	Previsto Negativo
Real Positivo	2556	671
Real Negativo	567	2660

Tabela 4.12: Matriz de Confusão - *Random Forest* - Teste Qui-Quadrado

	Previsto Positivo	Previsto Negativo
Real Positivo	2571	656
Real Negativo	585	2642

Neste teste devido à natureza do teste qui-quadrado, que não aceita valores negativos, foi necessário realizar a normalização das colunas. Feito isso, foi realizado o processo de *feature selection*, selecionando as 150 *features* com maior valor de dependência com as classes, por meio dos resultados do teste qui-quadrado. Após essa etapa, os modelos foram treiados e testados de acordo com o *10-fold cross validation*, e suas métricas foram obtidas. É possivel observar que os algoritmos tiveram desempenho similar, com o destaque do algoritmo XGBoost que apresentou as maiores médias das métricas de acurácia, *F1-Score*, precisão e *recall*, tal fato deu-se pela sua capacidade de combinar os múltiplos modelos empregando a técnica de *Gradient Boosting Tree*. Com desempenho similar o *Random Forest* ficou com as segundas maiores métricas e com o pior desempenho, apresentando os menores valores em comparação, ficou o *LightGBM*.

4.2.4 Discussão

Ao analisar os três testes o modelo que mais se destaca é o *XGBoost* que apresentou o melhor desempenho em relação aos modelos de *Random Forest* e *LightGBM*. O *XGBoost* obteve os maiores valores médios de acurácia, *F1-Score*, precisão e *recall* nos testes com os diferentes tipos de *feature selection*: *tree-based*, *Recusive Feature Eliminiation* e o teste qui-quadrado, destacando-se o método RFE o qual gerou os maiores

valores médios das métricas. O algoritmo *LightGBM* por sua vez teve bons resultados quando aplicado juntamente com a técnica *tree-based* e RFE, e o algoritmo *Random Forest* apresentou resultados próximos aos do *XGBoost* no teste em que foi utilizada a técnica de seleção de características por meio do teste qui-quadrado. Foi também realizado o teste ANOVA e os resultados obtidos comparando os 3 modelos levando em conta as diferentes técnicas de *feature selection* podem ser visualizadas na Tabela 4.13. A tabela é formada pelos seguintes campos: Amostra, que se refere aos modelos *LightGBM*, *XGBoost* e *Random Forest*, Colunas que representa as técnicas de *feature selection*, Interações que indica os efeitos combinados da Amostra e Colunas e Dentro representando a variação dentro de ambos os grupos. Na coluna SQ, é possivel visualizar os valores das soma dos quadrados, gl, os graus de liberdade e MQ a média dos quadrados. Ao analisar os valore de F, Valor-P e F-Crítico, é possível notar que existe um variância significativa no desempenho entre os modelos e as técnicas de *feature selection* utilizadas. Sendo assim, as interações entre os modelos e as técnicas de seleção de características podem variar os desempenhos obtidos de forma significativa.

Tabela 4.13: Tabela com os resultados de ANOVA para análise de variância dos modelos de *ensemble* e métodos de *feature selection*.

Fonte da Variação	**SQ**	**gl**	**MQ**	**F**	**Valor-P**	**F Crítico**
Amostra	0,000138587	2	$6{,}92936 \times 10^{-5}$	890,9178571	$2{,}23112 \times 10^{-25}$	3,354130829
Colunas	0,000568496	2	0,000284248	3654,614286	$1{,}37888 \times 10^{-33}$	3,354130829
Interações	0,000686553	4	0,000171638	2206,776786	$1{,}57431 \times 10^{-33}$	2,727765306
Dentro	$2{,}1 \times 10^{-6}$	27	$7{,}77778 \times 10^{-8}$			
Total	0,001395736	35				

Capítulo 5

Conclusão

A identificação dos exosítios de interação proteína-DNA e proteína-RNA é de suma importância para possibilitar o aprofundamento do conhecimento a respeito do funcionamento celular, o que pode possibilitar o desenvolvimento de novos fármacos e vacinas. Entretanto, os estudos na área são tradicionalmente realizados de forma experimental, sendo muito custosos e demandando muito tempo, sendo assim, métodos computacionais estão em ascensão. Este trabalho cumpriu o objetivo proposto ao utilizar algoritmos de aprendizado de máquina e *ensemble learning* para realizar a tarefa de classificação de exosítios que modulam as interações entre proteínas e ácidos nucleicos.

O processo teve início com o desenvolvimento de um banco de dados de resíduos da superfície de proteínas que fazem parte dos sítios de ligação e resíduos que não, devido às características do banco utilizado para realização da tarefa de classificação foi aplicado o *undersampling* a fim de balancear o banco, que então passou por uma etapa de *feature selection*, em que foram utilizados três métodos o *tree-based*, RFE e o teste qui-quadrado. Após essa etapa, os modelos *LightGBM*, *XGBoost* e *Random Forest* foram aplicados para classificação. Para avaliar os modelos, os testes seguiram a técnica de *10-fold cross-validation* e tiveram as métricas médias de acurácia, *F1-Score*, precisão e *recall*, juntamente com o desvio padrão, calculados. Essas métricas

foram utilizadas para a avaliação dos modelos ao comparar aqueles que tinham maiores médias em relação aos outros. O modelo *XGBoost* se destacou como o modelo com as maiores métricas quando aplicado juntamente com *feature selection tree-based*, *Recursive Feature Elimination* e o teste qui-quadrado, sendo que o teste utilizando o RFE gerou o melhor desempenho dentre os métodos de *feature selection*.

Como trabalho futuro, o uso de outros métodos de *feature selection* e aprendizado de máquina são possibilidades que podem ser exploradas. Além disso, o uso de redes neurais, a fim de realizar a classificação dos exosítios, também é uma alternativa, buscando melhor desempenho para a tarefa.

Referências Bibliográficas

AGGRAWAL, R.; PAL, S. **Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease**. *SN Computer Science*, Springer, v. 1, n. 6, p. 344, 2020.

ALSAHAF, A. et al. **A framework for feature selection through boosting**. *Expert Systems with Applications*, Elsevier, v. 187, p. 115895, 2022.

BELGIU, M.; DRĂGUŢ, L. **Random forest in remote sensing: A review of applications and future directions**. *ISPRS journal of photogrammetry and remote sensing*, Elsevier, v. 114, p. 24–31, 2016.

BOCK, P.; PANIZZI, P.; VERHAMME, I. **Exosites in the substrate specificity of blood coagulation reactions**. *Journal of Thrombosis and Haemostasis*, Wiley Online Library, v. 5, p. 81–94, 2007.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. **A review of feature selection methods on synthetic data**. *Knowledge and information systems*, Springer, v. 34, p. 483–519, 2013.

BOULESTEIX, A.-L. et al. **Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 2, n. 6, p. 493–507, 2012.

CHICCO, D.; TÖTSCH, N.; JURMAN, G. **The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation**. *BioData mining*, Springer, v. 14, p. 1–22, 2021.

CORSI, F. et al. **Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties**. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 16, n. 2, p. e1007624, 2020.

CUI, F. et al. **Protein–DNA/RNA interactions: Machine intelligence tools and approaches in the era of artificial intelligence and big data**. *Proteomics*, Wiley Online Library, v. 22, n. 8, p. 2100197, 2022.

DIJK, M. van; BONVIN, A. M. **A protein–DNA docking benchmark**. *Nucleic acids research*, Oxford University Press, v. 36, n. 14, p. e88–e88, 2008.

ERICKSON, B. J.; KITAMURA, F. ***Magician's corner: 9. Performance metrics for machine learning models***. [S.l.]: Radiological Society of North America, 2021. e200126 p.

FURLANETTO, G. C.; GOMES, V. Z.; BREVE, F. A. **Artificial Bee Colony Algorithm for Feature Selection in Fraud Detection Process**. In: SPRINGER. *International Conference on Computational Science and Its Applications*. [S.l.], 2023. p. 535–549.

GRANITTO, P. M. et al. **Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products**. *Chemometrics and intelligent laboratory systems*, Elsevier, v. 83, n. 2, p. 83–90, 2006.

JOHANSSON-ÅKHE, I.; MIRABELLO, C.; WALLNER, B. **Predicting protein-peptide interaction sites using distant protein complexes as structural templates**. *Scientific reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 4267, 2019.

KE, G. et al. **Lightgbm: A highly efficient gradient boosting decision tree**. *Advances in neural information processing systems*, v. 30, 2017.

LEE, W.; SEO, K. **Downsampling for binary classification with a highly imbalanced dataset using active learning**. *Big Data Research*, Elsevier, v. 28, p. 100314, 2022.

LI, Y.; GOLDING, G. B.; ILIE, L. **DELPHI: accurate deep ensemble model for protein interaction sites prediction**. *Bioinformatics*, Oxford University Press, v. 37, n. 7, p. 896–904, 2021.

MACHADO, M. R.; KARRAY, S.; SOUSA, I. T. D. **LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry**. In: IEEE. *2019 14th International Conference on Computer Science & Education (ICCSE)*. [S.l.], 2019. p. 1111–1116.

MAHAJAN, P. et al. **Ensemble learning for disease prediction: A review**. In: MDPI. *Healthcare*. [S.l.], 2023. v. 11, n. 12, p. 1808.

MIAO, J.; NIU, L. **A survey on feature selection**. *Procedia computer science*, Elsevier, v. 91, p. 919–926, 2016.

MIENYE, I. D.; SUN, Y. **A survey of ensemble learning: Concepts, algorithms, applications, and prospects**. *IEEE Access*, IEEE, v. 10, p. 99129–99149, 2022.

OBAHA, A.; NOVINEC, M. **Regulation of peptidase activity beyond the active site in human health and disease**. *International journal of molecular sciences*, MDPI, v. 24, n. 23, p. 17120, 2023.

OLIVEIRA, S. d. M. et al. **Sting_RDB: a relational database of structural parameters for protein analysis with support for data warehousing and data mining.** Genetics and Molecular Research, v. 6, n. 4, p. 911-922, 2007., 2007.

OSISANWO, F. et al. **Supervised machine learning algorithms: classification and comparison**. *International Journal of Computer Trends and Technology (IJCTT)*, v. 48, n. 3, p. 128–138, 2017.

PACHOULY, J. et al. **A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools**. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 111, p. 104773, 2022.

RANA, K. K. et al. **A survey on decision tree algorithm for classification**. *International journal of Engineering development and research*, IJEDR (www. ijedr. org), v. 2, n. 1, p. 1–5, 2014.

RODRÍGUEZ-LUMBRERAS, L. A. et al. **pyDockDNA: a new web server for energy-based protein-DNA docking and scoring**. *Frontiers in molecular biosciences*, Frontiers Media SA, v. 9, p. 988996, 2022.

SAGI, O.; ROKACH, L. **Ensemble learning: A survey**. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, Wiley Online Library, v. 8, n. 4, p. e1249, 2018.

SALIM, J. A. **Aplicação de técnicas de reconhecimento de padrões usando os descritores estruturais de proteínas da base de dados do software STING para discriminação do sítio catalítico de enzimas**. *Master's thesis. Campinas: UNICAMP, Faculdade de Engenharia Elétrica e Computação*, 2015.

SHARMA, H.; KUMAR, S. et al. **A survey on decision tree algorithms of classification in data mining**. *International Journal of Science and Research (IJSR)*, v. 5, n. 4, p. 2094–2097, 2016.

SHELKE, M. S.; DESHMUKH, P. R.; SHANDILYA, V. K. **A review on imbalanced data handling using undersampling and oversampling technique**. *Int. J. Recent Trends Eng. Res*, v. 3, n. 4, p. 444–449, 2017.

SHINDE, P. P.; SHAH, S. **A review of machine learning and deep learning applications**. In: IEEE. *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. [S.l.], 2018. p. 1–6.

SI, J.; ZHAO, R.; WU, R. **An overview of the prediction of protein DNA-binding sites**. *International journal of molecular sciences*, MDPI, v. 16, n. 3, p. 5194–5215, 2015.

SUKUMAR, S. et al. **DBSI server: DNA binding site identifier**. *Bioinformatics*, Oxford University Press, v. 32, n. 18, p. 2853–2855, 2016.

TANG, J.; ALELYANI, S.; LIU, H. **Feature selection for classification: A review**. *Data classification: Algorithms and applications*, CRC press, p. 37, 2014.

THASEEN, I. S.; KUMAR, C. A. **Intrusion detection model using fusion of chi-square feature selection and multi class SVM**. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, v. 29, n. 4, p. 462–472, 2017.

THASEEN, I. S.; KUMAR, C. A.; AHMAD, A. **Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers**. *Arabian Journal for Science and Engineering*, Springer, v. 44, p. 3357–3368, 2019.

WU, J. et al. **Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature**. *Bioinformatics*, Oxford University Press, v. 25, n. 1, p. 30–35, 2009.

XIONG, Y. et al. **Survey of computational approaches for prediction of DNA-binding residues on protein surfaces**. *Computational Systems Biology: Methods and Protocols*, Springer, p. 223–234, 2018.

YEGNESWARAN, S. et al. **Manipulation of thrombin exosite I, by ligand-directed covalent modification**. *Journal of Thrombosis and Haemostasis*, Elsevier, v. 5, n. 10, p. 2062–2069, 2007.

ZENG, M. et al. **Protein–protein interaction site prediction through combining local and global features with deep neural networks**. *Bioinformatics*, Oxford University Press, v. 36, n. 4, p. 1114–1120, 2020.