UNIVERSIDADE FEDERAL DE MINAS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

MARIA GABRIELA REIS CARVALHO

CONSTRUÇÃO DE MODELO PREDITIVO DE IMUNOGENICIDADE DE NEOEPÍTOPOS: aplicação de métodos de aprendizado estatístico no desenvolvimento de imunoterápicos contra o câncer

Belo Horizonte - MG
2023

MARIA GABRIELA REIS CARVALHO

CONSTRUÇÃO DE MODELO PREDITIVO DE IMUNOGENICIDADE DE NEOEPÍTOPOS: aplicação de métodos de aprendizado estatístico no desenvolvimento de imunoterápicos contra o câncer

Monografia apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais como requisito para recebimento do título de Especialista em Estatística

Orientador: Dr. Rafael Santos Erbisti (IME/UFF)

Belo Horizonte - MG

2023

Carvalho, Maria Gabriela Reis.

C331c Construção de modelo preditivo de imunogenicidade de neoepítopos: [recurso eletrônico] aplicação de métodos de aprendizado estatístico no desenvolvimento de imunoterápicos contra o câncer / Maria Gabriela Reis Carvalho. —2023.
1 recurso online (59 f. il, color.): pdf.

Orientador: Rafael Santos Erbisti.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 55-59

1. Estatística. 2. Câncer. 3. Imunoterapia. 4. Aprendizado Estatístico. 5.Aprendizado do computador. I. Santos, Rafael Erbist. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 289ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE MARIA GABRIELA REIS CARVALHO.

Aos dezesseis dias do mês de maio de 2023, às 14:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Maria Gabriela Reis Carvalho**, intitulado: "Construção de modelo preditivo de imunogenicidade de neoepítopos: Aplicação de métodos de aprendizado estatístico no desenvolvimento de imunoterápicos contra o câncer.", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Rafael Santos Erbisti – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada, incondicionalmente. As sugestões da banca deverão ser consideradas e a candidata deverá modificar o texto no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 16 de maio de 2023.

gov.br Documento assinado digitalmente
RAFAEL SANTOS ERBISTI
Data: 19/05/2023 12:09:37-0300
Verifique em https://validar.iti.gov.br

Prof.ª Rafael Santos Erbisti (Orientador)
Departamento de Estatística / IME / UFF

gov.br Documento assinado digitalmente
JESSICA QUINTANILHA KUBRUSLY
Data: 23/05/2023 08:58:37-0300
Verifique em https://validar.iti.gov.br

Prof.ª Jessica Quintanilha Kubrusly
Departamento de Estatística / IME / UFF

gov.br Documento assinado digitalmente
LUIZ CARLOS JUNIOR ALCANTARA
Data: 23/05/2023 10:43:02-0300
Verifique em https://validar.iti.gov.br

Luiz Carlos Júnior Alcântara
Instituto René Rachou / Fundação Oswaldo Cruz

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
P Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.

Declaro para os devidos fins que Maria Gabriela Reis Carvalho, número de registro 2020679994, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística e que entregou para seu orientador, o professor Rafael Santos Erbisti, o trabalho, que aprovou a versão final. O trabalho foi apresentado no dia 16 de Maio de 2023 com o título "Aplicação de métodos de aprendizado estatístico no desenvolvimento de imunoterápicos contra o câncer.".

Belo Horizonte, 27 de julho de 2023

Roberto da Costa Quinino:80871291720

Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2023.07.27 16:39:39 -03'00'

Prof. Roberto da Costa Quinino
Coordenador do curso de
Especialização em Estatística
Departamento de Estatística / UFMG

AGRADECIMENTOS

Com todo a força de meu coração, agradeço:

Ao Dr. Jeronimo Ruiz, líder do Grupo Informática de Biossistemas, Bioengenharia e Genômica da Fiocruz Minas, por me confiar a execução deste importante trabalho e acolher meu entusiasmo pueril.

Ao Professor Dr. Rafael Santos Erbisti, do Instituto de Matemática e Estatística da Universidade Federal Fluminense, por ter, com tamanha generosidade, dividido comigo os seus conhecimentos e por ter aceitado o desafio de me orientar com extrema paciência, respeito e bom humor. Rafael, você é nota 1.000! Espero, em algum momento, conseguir retribuir.

Ao Dr. Luiz Carlos Júnior Alcântara, pesquisador da Fiocruz Minas, e à Professora Dra. Jessica Quintanilha Kubrusly, do Instituto de Matemática e Estatística da Universidade Federal Fluminense (UFF), pela disponibilidade e interesse na leitura e avaliação deste trabalho, bem como pelas importantes discussões e contribuições.

Ao mestre e doutorando Paul Anderson, membro do Grupo Informática de Biossistemas, Bioengenharia e Genômica da Fiocruz Minas, meu grande parceiro neste trabalho. Obrigada pelas incontáveis horas, semanas e meses em que trabalhou ao meu lado, criando algoritmos que permitiram que quase todas as minhas ideias malucas de simulação fossem materializadas em tempo recorde. Paul, a você a minha admiração e profunda gratidão.

Ao meu esposo pela presença constante, incentivos, paciência, chás, biscoitos e frutas que refletiram o seu mais tenro cuidado, incentivo e amor. Não foi fácil, mas NÓS chegamos ao final.

À Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) pelo financiamento da *Rede Translacional de Imunoinformática: da identificação de biomarcadores ao desenvolvimento de estratégias imunoterapêuticas e microdispositivos para análises in vitro*" (RED-00104-22), na qual se inseriu este trabalho.

RESUMO

Este trabalho tem como objetivo principal contribuir com o desenvolvimento de novos imunoterápicos contra o câncer por meio do desenvolvimento e avaliação de modelos preditivos da capacidade de epítopos tumorais conduzirem à resposta imune efetiva. Um conjunto de 5.913 propriedades físico-químicas, estruturais e evolucionárias de sequências primárias de aminoácidos, consideradas potenciais preditoras da ativação das células TCD8+ foram empregadas na construção dos modelos. Modelos de regressão logística com penalização Lasso e *Group*-Lasso foram utilizadas para a seleção de um subconjunto ótimo de variáveis e os métodos de aprendizado estatístico XGBoost, Florestas Aleatórias (RF) e Máquinas de Vetores de Suporte (SVM) foram empregados para a classificação. Para o treinamento dos modelos, foi empregada a abordagem *bootstrap* com 1.000 amostras e, como métrica de desempenho, foram utilizados a Área Sob a Curva ROC (AUC), Sensibilidade e Especificidade. A análise dos modelos gerados permitiu verificar que, dentre os métodos de seleção de variáveis, o *Group*-Lasso foi aquele que apresentou melhor desempenho (AUC=0,6958). Dentre os modelos construídos com os diferentes métodos, a partir das variáveis selecionadas, aquele gerado com SVM apresentou um valor de AUC, igual a 0,8286, e Sensibilidade igual a 0,9078, tendo, portanto, um desempenho superior ao de modelos até então descritos na literatura.

Palavras-chave: epítopo; câncer; imunoterapia; aprendizado estatístico; aprendizado de máquinas.

ABSTRACT

The main objective of this work is to contribute to the development of new immunotherapies against cancer through the development and evaluation of predictive models of the ability of tumor epitopes to lead to an effective immune response. A set of 5,913 physical-chemical, structural and evolutionary properties of primary amino acid sequences, potentially predictive of CD8+ T cell activation, were used in the construction of the model. Logistic regression models with LASSO and Group-Lasso penalty were used for selection of an optimal set of variables and statistical learning methods XGBoost, Random Forests (RF) and Support Vector Machines (SVM) were used for classification. For training the models, the bootstrap approach was used with 1,000 runs and, as performance metrics, the Area under the ROC Curve (AUC), Sensitivity and Specificity were used. The analysis of the generated models allowed us to verify that, among the variable selection methods, the Lasso-Group presented the best performance (AUC=0.6958). Among the models built with the different methods, based on the selected variables, the one generated with SVM presented an AUC value equal to 0.8286 and Sensitivity equal to 0.9078, thus having a superior performance to the models presented so far in the literature.

Keywords: epitope; cancer; immunotherapy; statistical learning; machine learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

De acordo com a Organização Mundial de Saúde (OMS), o Câncer é a principal causa de morbidade e mortalidade no Mundo, com 19.292.789 novos casos e 9.958.133 mortes relacionadas à doença estimados para o ano de 2020. No Brasil, para o período entre 2023 e 2025, o número de novos casos estimados é de 704 mil. Na população brasileira, os tipos de cânceres mais frequentes são o de próstata (30% dos casos) na população masculina e o câncer de mama (30,1% dos casos) na população feminina. Tanto para os homens quanto para as mulheres, o câncer colorretal é aquele que figura na segunda posição, com uma frequência, em 2022, de 9,2-9,7% dos casos de câncer no país (“Estimativa 2023: incidência de câncer no Brasil | INCA - Instituto Nacional de Câncer”, [s.d.]).

Para o tratamento do câncer, diferentes abordagens podem ser empregadas, tais como cirurgia, quimioterapia, radioterapia, terapia hormonal e a imunoterapia. A escolha da abordagem terapêutica depende de diferentes fatores como, por exemplo, tipo do tumor, estágio de desenvolvimento e características clínicas do indivíduo acometido pela doença. No Brasil, o tratamento para o câncer está disponível na rede do Sistema Único de Saúde e é ofertado nas Unidades de Assistência de Alta Complexidade em Oncologia ou Centros de Assistência de Alta Complexidade em Oncologia, conforme a Política Nacional de Prevenção e Controle do Câncer (BRASIL, 2013).

Considerando-se as abordagens terapêuticas atualmente disponíveis para o tratamento do câncer, se verifica que desde a última década a imunoterapia vem sendo considerada a mais promissora e revolucionária, em função de seu potencial curativo, especificidade e efeitos colaterais potencialmente mais brandos quando se comparado às terapias convencionais (WALDMAN; FRITZ; LENARDO, 2020). Na imunoterapia, o próprio sistema imunológico do paciente é utilizado como ferramenta para combater o tumor (ZHANG; ZHANG, 2020). Para isso, diversos métodos podem ser empregados, tais como os inibidores de *checkpoint* imunológico, vacinas terapêuticas e transferência de células T.

Embora os métodos empregados para o tratamento imunoterapêutico sejam diversos, em geral, muitos deles envolvem a ativação dos linfócitos TCD8+ a partir do reconhecimento de fragmentos de antígenos tumorais, os chamados neoepítopos. Os neoepítopos são peptídeos que, quando ligados às moléculas de MHC I, são apresentados às células do sistema imune. Verifica-se que, no organismo humano, os epítopos tumorais são naturalmente capazes de suscitar uma resposta imunológica. Esta resposta, entretanto, é ineficiente e incapaz de combater o tumor. Assim, o fundamento de grande parte dos métodos de imunoterapia é

potencializar o reconhecimento destes epítopos pelo sistema imune, gerando uma resposta mais eficaz (WALDMAN; FRITZ; LENARDO, 2020).

A identificação de neoepítopos capazes de induzir respostas imunes, entretanto, é um processo laborioso. Os métodos experimentais demandam grande número de amostras e análises, o que está associado a elevados custos e dispêndio de tempo. Neste sentido, as abordagens computacionais são uma alternativa importante para otimização do processo de identificação de neoepítopos imunogênicos.

Os modelos desenvolvidos para a predição de epítopos tumorais, em sua maioria, estão voltadas ao processamento dos antígenos em meio intracelular e à sua ligação à molécula de MHC I. Embora estas etapas sejam extremamente relevantes, evidências indicam que a maior parte dos neoepítopos preditos não são de fato imunogênicos (CAI et al., 2022; FOTAKIS; TRAJANOSKI; RIEDER, 2021). Assim, é necessário avanços em modelos preditivos que envolvam também o reconhecimento do complexo peptídeo-MHC pelas células TCD8+ e, neste sentido, os métodos de aprendizado estatístico são uma estratégia interessante para auxílio à tomada de decisões a partir de grandes conjuntos de dados.

Estudos como os de Tung e Ho (2007, 2011), Saethang et al. (2013) e Zhang et al. (2015) propuseram modelos preditivos de imunogenicidade de epítopos, utilizando métodos de aprendizado estatístico, e construídos a partir de diferentes propriedades físico-químicas dos aminoácidos que formam a cadeia peptídica do epítopo. Observa-se, entretanto, que embora estes métodos tenham alcançado bons valores de medidas de desempenho preditivo, verifica-se que foram treinados e validados com bancos de dados que não são específicos de tumores, o que compromete a sua translação para os epítopos tumorais.

Uma vez que poucos trabalhos, até o momento, se voltam ao desenvolvimento e avaliação de modelos preditivos de imunogenicidade de neoepítopos tumorais (DIAO et al., 2022; LIU; SHI; LI, 2020; XIE; SHI; ZHANG, 2021) e dada a relevância deste tipo de predição para o desenvolvimento de novos imunoterápicos que possam contribuir com avanços no tratamento de diferentes tipos de câncer, este estudo se volta ao processo de construção de um novo modelo preditivo empregando diferentes métodos de aprendizado estatístico.

2 OBJETIVOS

2.1 Objetivo Geral

Contribuir com o desenvolvimento de novos imunoterápicos contra o câncer a partir do desenvolvimento de modelo preditivo de imunogenicidade de epítopos tumorais baseado em estratégias de Aprendizado Estatístico.

2.2 Objetivos Específicos

Avaliar o desempenho preditivo de modelos de imunogenicidade utilizando diferentes métodos de regularização, a partir da seleção de um subconjunto de variáveis estruturais, físico-químicas e evolucionárias de epítopos imunogênicos e não imunogênicos.

Construir modelos preditivos de imunogenicidade de epítopos utilizando diferentes métodos de Aprendizado de Máquina (Florestas Aleatórias, Máquinas de Vetores de Suporte e XGboost), identificando aquele com melhor desempenho para a predição da imunogenicidade de epítopos.

Avaliar o impacto da utilização das variáveis selecionadas por meio de regressão logística associada a métodos de regularização na capacidade preditiva dos modelos construídos com as estratégias de Aprendizado de Máquinas.

Construir um modelo preditivo de imunogenicidade de epítopos a partir da combinação de diferentes métodos de Aprendizado de Máquina, avaliando o seu desempenho em relação àquele apresentado pelos modelos individuais.

3 REFERENCIAL TEÓRICO

3.1 Câncer

O Câncer[1] é uma doença genética complexa caracterizada pelo crescimento descontrolado de células anormais que tendem invadir outros tecidos. Todas as diferentes populações celulares do organismo humano podem ser acometidas pelas alterações que caracterizam o câncer, o que faz que este termo se refira a um espectro de mais de 100 doenças (BUNZ, 2022).

Os traços fundamentais compartilhados pela maioria dos tumores podem ser classificados nas seguintes categorias: “capacidades funcionais” ou “características habilitantes” (HANAHAN, 2022). As capacidades funcionais são características adquiridas pelas células sadias à medida que sofrem o processo de transformação para células neoplásicas. São elas:

a) autossuficiência na sinalização relacionada à proliferação, de modo que as células tumorais não dependem exclusivamente de estímulos externos;

b) redução da sensibilidade aos sinais antiproliferativos que mantém a homeostase tecidual;

c) capacidade para evadir do processo de morte programada (apoptose), o que faz com que as células cancerosas possam se tornar imortais e, se mantendo sempre vivas, continuando a se proliferar;

d) potencial proliferativo ilimitado;

e) invasão tecidual e metástase, o que está relacionado à produção, pelas células tumorais, de proteínas que modulam sua relação com o ambiente circundante;

f) instabilidade do genoma e aquisição de múltiplas mutações, o que conduz às outras diferentes características apresentadas pelas células tumorais;

g) reprogramação do metabolismo energético, o que está relacionado, por exemplo, à utilização, pelas células tumorais, de vias glicolíticas pouco utilizadas pelas células sadias e que são capazes gerar, mais rapidamente aminoácidos que sustentam o seu processo de proliferação.

[1] Para este trabalho, serão considerados como sinônimos para câncer, as seguintes palavras: tumor e neoplasia.

As características habilitantes, por outro lado, são fatores que permitem que as células sadias adquiriam as capacidades funcionais que as tornam neoplásicas. Formam este grupo de características:

a) evasão do sistema imune, uma vez que mecanismos de escape permitem que as células tumorais não sejam identificadas como componentes não próprios do organismo;

b) inflamação promotora do tumor, já que mecanismos inflamatórios teciduais desencadeados pela presença das células cancerosas contribuem com a sua proliferação, sobrevivência, formação de vasos para nutrição tumoral e produção de moléculas capazes de promover mutações em células adjacentes;

c) desbloqueio da plasticidade fenotípica, o que faz com que as células tumorais adquiriam capacidade de diferenciar em outros tipos celulares;

d) reprogramação epigenética, relacionada à modulação do tumor pelo microambiente;

e) células senescentes adjacentes que podem produzir moléculas sinalizadoras que auxiliam a modular as características funcionais do câncer;

f) microbiomas polimórficos, envolvendo, por exemplo, bactérias que são capazes de mimetizar sinais proliferativos que conduzem à expansão do câncer.

3.1.1 Genética do Câncer

Os processos neoplásicos, que estão relacionados à aquisição, pelas células saudáveis, de competências funcionais específicas do câncer, estão relacionados a alterações genéticas (GERDES, 2002). Os genes são trechos das moléculas de DNA[2], formados por uma sequência de nucleotídeos, que codificam os processos biológicos do organismo, como a produção de proteínas. Até o momento, cerca de 500 diferentes genes já foram identificados como fortemente relacionados ao processo de transformação das células saudáveis em células tumorais (UHLEN et al., 2017).

No organismo, as proteínas codificadas pelos genes têm papel crítico tanto para a formação da estrutura das células e tecidos, quanto para a transdução de sinais e controle de

[2] DNA (ácido desoxirribonucleico) é uma molécula complexa, responsável pela codificação dos processos biológicos do organismo. A estrutura do DNA é formada por longas cadeias poliméricas duplas, compostas por diferentes ácidos nucléicos, ou nucleotídeos: adenina (A), citosina (C), guanina (G) e timina (T), dispostos em uma determinada ordem, formando uma sequência.

importantes funções que estão alteradas no câncer, como a proliferação, renovação, morte e movimentação celular. Cada tipo de proteína é formado por uma determinada sequência de aminoácidos[3] e gerada por meio de um processo de produção que envolve: a transcrição dos genes em moléculas de RNA mensageiro (mRNA); tradução dos nucleotídeos presentes no mRNA em aminoácidos específicos; e formação da cadeia polipeptídica da proteína por meio da ligação entre os aminoácidos, de acordo com a ordem da sequência de nucleotídeos do gene codificante(ALBERTS et al., 2017).

3.1.1.1 Mutações no Genoma Associadas ao Câncer

No câncer, mutações nas sequências de nucleotídeos que definem a estrutura e a função dos genes, derivadas de danos no DNA causados por exemplo, por exposição a fatores ambientais como tabaco e raios UV, infecções virais e erros durante o processo de divisão celular(BUNZ, 2022), conduzem a modificações da expressão, estrutura e função das proteínas formadas. As mutações mais associadas à doença são as substituições de nucleotídeo único (SNVs) e inserções e deleções (INDEL).

As SVNs, também chamadas de mutações pontuais, são aquelas em que há alteração de um único par de base na sequência do DNA. Elas podem ser: a) sinônimas, quando a mutação do nucleotídeo não compromete a formação do aminoácido; b) *missense*, quando a mutação altera o aminoácido formado, podendo resultar em alteração da estrutura e função da proteína; c) *nonsense*, quando um *stopcodon* é gerado na sequência do gene, interrompendo a formação da proteína e gerando uma molécula não funcional.

As mutações INDEL são aquelas nas quais um par de base ou um conjunto de pares de bases são deletados ou inseridos na sequência do DNA. Quando o número de nucleotídeos inseridos ou deletados é múltiplo de 3 pares de bases, a proteína mantém sua sequência inalterada. Já quando o número de nucleotídeos não é múltiplo de 3, pode haver uma parada prematura da síntese da proteína ou seu alongamento.

Outro tipo de mutação gênica verificada nos cânceres são as modificações estruturais. Nestes casos, se verifica alterações de mais de 50 pares de bases nos cromossomos, podendo resultar em várias cópias de um mesmo gene no DNA. Esta alteração resulta em níveis mais elevados da proteína codificada(BUNZ, 2022).

De um modo geral, os genes afetados por estas mutações (cerca de 1% do genoma humano), são: os proto-oncogenes e os genes supressores de tumor. Os proto-oncogenes

[3] Existem, no corpo humano, 20 tipos diferentes de aminoácidos que possuem uma estrutura básica comum e um grupo lateral responsável pelas características químicas específicas de cada um dos tipos.

são genes que regulam os processos celulares básicos relacionados ao crescimento e a diferenciação celular. Mutações nestes genes podem conduzir à sua transformação em oncogenes que induzem uma superprodução de proteínas ou a produção de proteínas com maiores níveis de atividade bioquímica que àquelas apresentadas pelas proteínas codificadas pelos proto-oncogenes. Esta desregulação resulta em aumento descontrolado da proliferação celular, característico do câncer. Os genes supressores de tumor, por outro lado, têm como função manter a estabilidade dos tecidos, por meio da regulação da progressão do ciclo celular, diferenciação das células e do processo de morte programada (apoptose), além da manutenção da integridade genética. Mutações que induzem a inativação dos genes supressores de tumor resultam na perda da homeostase, crescimento e progressão tumoral (ALBERTS et al., 2017; BUNZ, 2022).

É interessante observar que as células do organismo humano estão constantemente sujeitas a danos no DNA em função de exposição a fatores exógenos, por exemplo. Estes danos, entretanto, não implicam, necessariamente, no desenvolvimento de um câncer. Isso acontece porque, para este desenvolvimento, é necessário um acúmulo de mutações sucessivas em uma célula. Além disso, proteínas específicas, codificadas pelos genes supressores de tumor, são capazes de interromper o ciclo celular até que o DNA seja reparado ou conduzir as células que sofreram mutações a um processo de morte programada (BUNZ, 2022).

Outro elemento que pode estar associado à capacidade do organismo em conter a progressão tumoral é o sistema imune. Conforme abordado anteriormente, as alterações genéticas associadas ao surgimento do câncer e à sua progressão estão diretamente relacionadas à produção de proteínas anormais pelas células cancerosas. Estas proteínas são expressas pelas células cancerosas e podem ativar o sistema imune, estimulando a destruição das células alteradas antes do desenvolvimento do tumor (vigilância imunológica). Estas respostas imunológicas, embora existentes, muitas vezes não são efetivas. Tais aspectos conduzem à ideia de desenvolvimento de novas estratégias de tratamento do câncer capazes de potencializar as respostas imunológicas anti-tumorais.

3.1.1.2 Respostas Imunes e o Câncer

O sistema imune se refere a um conjunto de células, tecidos e moléculas envolvidos nos mecanismos de defesa contra agentes reconhecidos como estranhos pelo organismo, como patógenos (vírus, bactérias, fungos, protozoários), tumores, toxinas e até mesmo componentes próprios. O processo de reação do sistema imune a estes agentes é

denominado resposta imune e pode ser entendido como duas grandes linhas de defesa: resposta inata e resposta adaptativa (ABUL K. ABBAS; SHIV PILLAI, 2019).

A primeira linha de defesa, a imunidade inata (imunidade natural ou imunidade nativa) é uma resposta não especializada, consistindo em mecanismos celulares e bioquímicos que são desencadeados imediatamente após a exposição ao agente ou ainda nas primeiras horas de infecção. A segunda linha de defesa é a imunidade adaptativa, um tipo de resposta específica. Neste tipo de resposta, as substâncias solúveis, celulares ou particuladas, denominadas de antígenos, são reconhecidas por receptores de linfócitos T e por receptores e moléculas que são secretadas pelos linfócitos B (ABBAS; LICHTMAN; PILLAI, 2015).

A especificidade da resposta adaptativa é possível porque clones de linfócitos antígenos-específicos, chamados linfócitos *naive*, se desenvolvem antes e de modo independente da exposição do organismo ao antígeno. Cada clone possui um único receptor de antigênico, ou seja, só se liga a um tipo específico de antígeno. Estes clones circulam pelo organismo e, quando encontram um antígeno compatível com sua estrutura, eles se ligam. Esta ligação, por sua vez, desencadeia a proliferação de clones e sua diferenciação em células com a mesma especificidade: a) células efetoras, capazes de destruir o antígeno; b) células de memória, capazes de reagir vigorosamente, caso a infecção se prolongue ou persista. A repetição da exposição do organismo ao antígeno aumenta sua capacidade de reposta, uma vez que cada exposição resulta na geração de novas células de memória que têm a capacidade de reagir mais rapidamente e de modo mais ampliado ao estímulo que os clones *naive* (ABBAS; LICHTMAN; PILLAI, 2015).

No câncer, as respostas imunes clinicamente relevantes são as repostas adaptativas envolvendo o reconhecimento de antígenos tumorais pelos linfócitos T. Uma vez que este tipo de resposta pode limitar o crescimento e a disseminação dos tumores, sendo relevante no contexto da imunoterapia, ela será o foco do presente tópico.

Os antígenos tumorais são proteínas produzidas durante a instalação e desenvolvimento do câncer. Para que sejam reconhecidos pelas células T, os antígenos tumorais são quebrados, no interior das células, em pequenos peptídeos (epítopos) e devem se ligar a moléculas do *Major Histocompatibily Complex* (MHC)[4], formando um complexo MHC-peptídeo

[4] As moléculas de MHC são glicoproteínas transmembranas, que possuem uma fenda para a ligação do peptídeo e que são capazes de apresentar, no meio externo às células, os antígenos a elas ligados, possibilitando o reconhecimento do epítopo pelas células T.

que é apresentado na superfície celular. Apenas peptídeos apresentados pelo MHC são reconhecidos pelas células TCD4+ e TCD8+[5], desencadeando a resposta imune.

Os antígenos tumorais podem ser associados ao tumor (TAA) ou tumor específicos (TSA). Os TAAs são superexpressos pelas células cancerosas, mas também são expressos nos tecidos saudáveis. Uma vez que fazem parte do sistema imunológico normal, não são capazes de desencadear respostas protetoras efetivas. Os TSA, também conhecidos como neoantígenos, por outro lado, são derivados de proteínas mutadas produzidas exclusivamente pelas células tumorais. Em função da especificidade, as células T capazes de reconhecer estes neoantígenos podem escapar do mecanismo de seleção negativa, conduzindo a geração de uma resposta imune mais eficiente (SMITH et al., 2019).

Os neoepítopos são apresentados na superfície das células tumorais pelas moléculas de MHC I (Figura 1). Este complexo neoepítopo-MHC, quando reconhecido pelos receptores presentes nas células TCD8+ (citotóxicas)[6], desencadeia a resposta protetora.

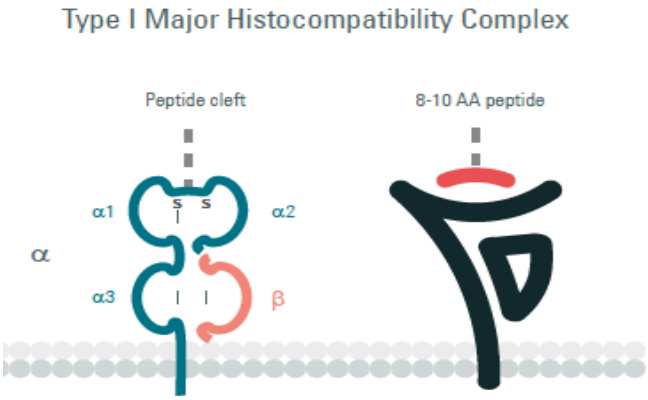

Figura 1- Representação esquemática da molécula de MHC.
À esquerda, verifica-se a molécula, com suas diferentes cadeias e, à direita, a molécula com um peptídeo ancorado, indicando que parte dele está ligado ao MHC e outra parte é livre para reconhecimento e ligação aos linfócitos TCD8+.

Conforme apresentado na Figura 1, as moléculas de MHC I são formadas por duas cadeias peptídicas ligadas por meio de ligação não covalente: a) cadeia pesada, composta por de 3 domínios α microgobulina (α1, α2 e α3); b) cadeia leve, composto por uma β microgobulina. A molécula de MHC I fica ancorada à membrana, por meio de uma pequena

[5] Linfócitos TCD4+ são os chamados “Linfócitos citotóxicos” e Linfócitos TCD8+ são os chamados linfócitos *helper*.
[6] Wells et al., “Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction”.

porção hidrofóbica de sua cadeia pesada, e sua maior parte se projeta para fora do citoplasma da célula, formando uma cavidade antigênica com as α1 e α2 microglobuilnas (na qual se liga um peptídeo composto por 8 a 10 aminoácidos) e uma alça (α3 e β microglobulinas), que serve de ligação para as células TCD8+.

3.2 Desenvolvimento de Imunoterápicos e Imunogenicidade de epítopos

Conforme discutido nos tópicos anteriores, o câncer se trata de uma doença genética decorrente de mutações no DNA que afetam genes que codificam proteínas envolvidas na regulação de importantes atividades celulares, como a proliferação e apoptose. A transformação das células saudáveis em células tumorais, sua proliferação anormal, bem como a invasão de outros tecidos estão relacionados ao comprometimento do funcionamento de tecidos e órgãos.

Apesar das opções terapêuticas existentes para tratamento do câncer, como cirurgia, quimioterapia, terapia hormonal e radioterapia, as elevadas taxas de morbidade e mortalidade associadas à doença conduzem à intensificação das pesquisas e desenvolvimentos tecnológicos relacionados a novas abordagens. Neste cenário de novos desenvolvimentos, a imunoterapia, envolvendo como ferramenta a utilização de componentes do sistema imune, vem sendo considerada uma das mais promissoras e poderosas estratégias para tratamento do câncer em função de seu maior potencial curativo e efeitos colaterais menos agressivos em relação às terapias convencionais (HU; OTT; WU, 2018).

A utilização da imunoterapia para tratamento do câncer está baseada em evidências que demonstram a teoria de que o sistema imune funciona em constante vigilância e tem capacidade para reconhecer e eliminar as células anormais (imunovigilância), principalmente pela ação dos linfócitos TCD8+. Este mecanismo, entretanto, é ineficiente e as células tumorais podem evadir dos mecanismos de reconhecimento e defesa, desencadeando a instalação e progressão do tumor (RIBATTI, 2017). Assim, o foco da imunoterapia é ativar ou potencializar as respostas do próprio sistema imune de modo que ele seja capaz de atacar, de modo eficiente, as células tumorais por meio de seus mecanismos naturais (RILEY et al., 2019)

Diferentes classes de imunoterápicos vêm sendo desenvolvidos para o tratamento do câncer, as quais incluem, mas não se limitam aos bloqueadores de *checkpoint* imunológico, que são desenvolvidos para desencadear respostas mais poderosas pelas células TCD8; as terapias de células adotivas, que estão relacionadas à infusão, no organismo, de células

imunológicas; e vacinas terapêuticas, que ensinam o sistema imune a reconhecer e destruir as células tumorais (RILEY et al., 2019). Estas classes de imunoterápicos, apesar de distintas tecnicamente, estão associadas ativação das células T desencadeada pelo reconhecimento de epítopos tumorais (WALDMAN; FRITZ; LENARDO, 2020). Deste modo, para o desenvolvimento de novas moléculas, a identificação de epítopos capazes de serem reconhecidos pelos linfócitos T e desencadearem resposta imune adaptativa efetiva é, portanto, um processo central (HU; OTT; WU, 2018; LIU et al., 2022).

Evidências indicam que os epítopos específicos do tumor, denominados neoepítopos, quando comparados aos epítopos associados ao tumor, são mais efeitos na geração de uma resposta antitumoral pelas células TCD8+ e tem maior potencial para o desenvolvimento de imunoterápicos, uma vez que não são expressos pelas células sadias (JIANG et al., 2019; XIE et al., 2023). A identificação experimental de neoepítopos capazes de elicitar uma resposta imune efetiva, entretanto, é um processo longo, caro e laborioso (GARCIA-GARIJO; FAJARDO; GROS, 2019). Neste cenário, o desenvolvimento de novos imunoterápicos vem se beneficiando da utilização das tecnologias de sequenciamento de nova geração e de predições computacionais para a otimização do processo de identificação de peptídeos tumorais capazes de induzir a resposta imune (XIE et al., 2023).

Conforme discutido anteriormente, para que um epítopo seja apresentado às células do sistema imune e possa desencadear uma resposta pelas células TCD8+, é necessário, primariamente, que ele esteja ligado a uma molécula de MHC. Assim, em geral, as abordagens computacionais para a predição de neoepítopos estão voltadas ao processamento dos antígenos em meio intracelular, como clivagem, transporte do peptídeo e ligação à molécula de MHC (CAI et al., 2022; FOTAKIS; TRAJANOSKI; RIEDER, 2021) .

Um ponto importante, entretanto, é que a existência do peptídeo e sua ligação com a molécula de MHC não necessariamente indica que ele seja capaz de induzir resposta imune. É necessário ainda que as células TCD8+ reconheçam aquele peptídeo como não próprio e que o complexo peptídeo-MHC se ligue ao receptor da célula T. De fato, a maior parte dos peptídeos candidatos, que se ligam às moléculas de MHC I não são reconhecidos pelas células T e não são capazes de desencadear uma resposta imune efetiva (FOTAKIS; TRAJANOSKI; RIEDER, 2021). A etapa de análise do reconhecimento do complexo peptídeo-MHC pelas células TCD8+ e a deflagração das respostas imunes, por outro lado, tradicionalmente é realizada por meio de experimentos em bancada (CAI et al., 2022; FOTAKIS; TRAJANOSKI; RIEDER, 2021)

Considerando-se a necessidade de avançar em predições relacionadas não apenas à ligação dos peptídeos às moléculas de MHC I, novos métodos computacionais foram desenvolvidos incorporando a ideia de imunogenicidade. Um dos primeiros algoritmos foi o POPI (TUNG; HO, 2007), construído a partir de um conjunto de 531 propriedades fisico-químicas como variáveis preditoras da ligação entre o MHC I e os receptores de células T. O algoritmo inclui a mineração de variáveis para a seleção das mais importantes para a construção do modelo de classificação utilizando SVM. Seu treinamento e teste foi realizado com o conjunto de dados PEPMHC I, com diversos peptídeos associados às moléculas de MHC I e resultou em um modelo final com 23 variáveis e acurácia de 64,72%.

Uma variação do algoritmo POPI é o POPSKI (TUNG et al., 2011), que incorporou, entre as variáveis preditoras, informações relacionadas à posição dos aminoácidos na cadeia peptídica. Isso porque a posição dos aminoácidos poderia agregar informações sobre a estrutura do peptídeo e sobre sítios de ligação às moléculas de MHC I. Assim como o POPI, o modelo de classificação do POPSKI foi construído com o método de Máquinas de Vetores de Suporte[7]. Os processos de treinamento e validação do modelo foram realizados com o banco de dados IMMA2, formado por 1.085 epítopos imunogênicos e não imunogênicos. O valor de acurácia obtido foi de 68%.

Recentemente, Zhang et al. (2015) propuseram um novo modelo de predição de imunogenicidade de epítopos que avançou em relação àqueles desenvolvidos anteriormente. Os autores empregaram algoritmo genético para seleção de um conjunto ótimo de variáveis a partir de um *conjunto de dados* que incorporava diferentes categorias de variáveis (físicas, químicas e estruturais) que potencialmente poderiam auxiliar na predição de imunogenicidade de epítopos. Para a classificação, foi empregado o método de Florestas Aleatórias. O modelo alcançou um desempenho igual a 0,846 (Área sob a Curva ROC) para um *conjunto de dados* de treino.

Apesar dos grandes avanços apresentados em trabalhos como os de Tung e HO (2007), Tung et al. (2011) e Zhang et al (2015), é importante destacar que tanto o treino quanto a validação dos modelos preditivos não foram realizadas com bancos de dados de amostras tumorais. Uma vez que as mutações que conduzem ao câncer, em geral, são pontuais, a predição de imunogenicidade de epítopos tumorais é mais desafiadora e os métodos atuais não se mostram efetivos. Neste sentido, novos estudos voltados ao desenvolvimento de novos modelos preditivos que auxiliem o desenvolvimento de novos imunoterápicos se fazem

[7] Os métodos de aprendizado estatístico serão detalhados na subseção 3.3.

necessários. Neste contexto de desenvolvimento, as abordagens de aprendizado estatístico têm um importante papel.

3.3 Aprendizado Estatístico

O aprendizado estatístico se refere a um conjunto de diferentes métodos estatísticos que são empregados para construir modelos que dão sentido a complexos conjuntos de dados (JAMES et al., 2013). Um dos importantes métodos do aprendizado estatístico envolve a classificação, cujo objetivo é, a partir de uma amostra com observações $(X_1, Y_1), \ldots, (X_n, Y_n)$, sendo $\boldsymbol{X}$ um vetor com variáveis independentes e $\boldsymbol{Y}$ um vetor com as variáveis respostas qualitativas, construir uma função de predição $\mathrm{g}(x)$ que possa ser utilizada para predizer a classe das variáveis respostas de novas observações a partir dos valores das suas variáveis independentes. Os modelos preditivos que permitem a classificação dos dados vêm sendo empregados em diferentes trabalhos relacionados ao câncer, como a predição do diagnóstico, prognóstico, novas moléculas terapêuticas, respostas a diferentes tratamentos e desfechos clínicos em câncer (KOUROU et al., 2021).

Dentre os métodos de aprendizado estatístico voltados à classificação estão a Regressão Logística, Máquina de Vetores de Suporte, Árvores de Classificação, Florestas Aleatórias, métodos *Boosting*, métodos *Stack*, Redes Neurais Artificiais e *K*-Vizinhos Mais Próximos. A seguir, serão apresentados os modelos logístico, Máquina de Vetores de Suporte, Árvores de Classificação, Florestas Aleatórias, métodos *Boosting* e métodos *Stack*, que foram selecionados para a construção dos modelos de predição de imunogenicidade de epítopos proposta neste trabalho.

3.3.1 Modelo Logístico

Modelos de regressão são aqueles que permitem determinar a relação entre uma variável resposta aleatória Y e um vetor de variáveis respostas $\boldsymbol{x} = (x_1, \ldots, x_d) \in R^d$, estimando uma função de regressão

$$r_x = E[Y|\boldsymbol{X} = \boldsymbol{x}]$$

O modelo logístico é um dos principais métodos aplicados para classificação ou predição quando a variável resposta $Y_i,\ i = 1, \ldots, n$, assume apenas dois valores possíveis sendo, portanto, dicotômica ou binária (JAMES et al., 2013). Para isso, se assume que a variável

resposta tem distribuição Bernoulli ($Y_i \sim Bernoulli(p_i)$), sendo p_i a probabilidade de sucesso. O modelo de regressão logística pode ser representado a partir das equações abaixo

$$Y_i \sim Bernoulli(p_i)$$

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Uma vez que $0 < p_i < 1$, a função de probabilidade pode ser dada por:

$$E(Y_i) = p_i = P(Y_i = 1|\boldsymbol{x}) = \frac{e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}{1 + e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}$$

$$1 - p_i = P(Y_i = 0|\boldsymbol{x}) = \frac{1}{1 + e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}$$

E, portanto:

$$\frac{p_i}{1-p_i} = e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}$$

Sendo $\frac{p_i}{1-p_i}$, chamado *Odds* (ou chance, em português), que pode ser entendido como a razão entre a probabilidade de ocorrência sobre a probabilidade de não ocorrência do evento de interesse. Para interpretar os coeficientes de regressão, é utilizada a razão de chances (*Odds Ratio*).

Para estimar os parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ do modelo de regressão logística, é utilizada a função de verossimilhança:

$$\mathrm{L}(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^{p} \left(\frac{e^{(x_i^T \boldsymbol{\beta})}}{1 + e^{(x_i^T \boldsymbol{\beta})}}\right)^{y_i} \left(\frac{e^{(x_i^T \boldsymbol{\beta})}}{1 + e^{(x_i^T \boldsymbol{\beta})}}\right)^{1-y_i}$$

$$= \mathrm{L}(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^{p} \left(\frac{e^{(x_i^T \boldsymbol{\beta})^{y_i}}}{1 + e^{(x_i^T \boldsymbol{\beta})}}\right)$$

Os parâmetros de $\beta_0, \beta_1, \dots, \beta_p$ selecionados são aqueles para os quais a função de verossimilhança atinge um valor máximo. Assim, os estimadores de máxima verossimilhança

são os valores $\beta_0, \beta_1, \ldots, \beta_p$ que maximizam a função de verossimilhança. A função objetivo é dada por:

$$max_{\boldsymbol{\beta}} \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{(x_i^T \boldsymbol{\beta})^{y_i}}}{1 + \mathrm{e}^{(x_i^T \boldsymbol{\beta})}} \right)$$

Ou, de modo semelhante:

$$min_{\boldsymbol{\beta}} \frac{-1}{n} \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) - log \left(1 + exp \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) \right) \right]$$

3.3.2 Penalização Lasso

O viés de um modelo se refere à diferença entre a estimativa média do modelo e o valor real que se está tentando prever. Já a variância captura o quanto as estimativas do modelo se alteram quando o treino é feito com um *conjunto de dados* diferente. Quando o viés de um modelo é elevado, entende-se que ele é muito simplificado e apresentará um desempenho ruim na etapa de teste (sub-ajuste ou *underfitting*). Modelos com alta variância indicam que eles foram bem ajustados a um determinado conjunto de dados, mas que falharão nas predições realizadas com o conjunto de teste (super-ajuste ou *overfitting*). Assim, a capacidade de generalização do modelo é reduzida. Neste contexto, verifica-se que a capacidade preditiva do modelo está associada a um balanço entre o viés e a variância (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Problemas envolvendo a geração de modelos preditivos a partir de *conjunto de dados* com alta dimensionalidade, ou seja, aqueles em que o número de covariáveis é superior ao número de observações, geralmente se observa elevada variância, havendo, portanto, um *overfitting* e maior risco estimado associado. Nestes casos, a fim de reduzir a variância dos estimadores e melhorar a capacidade preditiva do modelo, uma estratégia possível é inserir viés nas estimativas, obtendo-se um preditor mais parcimonioso. Isso pode ser feito por meio da implementação de métodos de regularização ou *Shrinkage* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Um dos métodos de regularização é o *Last Absolute Shrinkage and Selection Operator* (LASSO) (TIBSHIRANI, 1996). No caso dos modelos logísticos, o LASSO adiciona uma penalização ao termo negativo do logaritmo da função de máxima de verossimilhança, minimizando os valores de $\boldsymbol{\beta}$.

$$\boldsymbol{\beta}^{lasso} = argmin_{\boldsymbol{\beta}} \left\{ \frac{-1}{n} \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) - log \left(1 + exp \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) \right) \right] + \lambda ||\boldsymbol{\beta}||_1 \right\}$$

No termo $\lambda||\boldsymbol{\beta}||_1$, verifica-se que a penalização ou minimização dos parâmetros $\boldsymbol{\beta}$ é modulada pelo parâmetro λ (*tuning parameter*). Quanto menor o valor de λ mais variáveis entram no modelo. Para $\lambda = 0$, o modelo se comporta como uma regressão logística usual, utilizando estimadores de máxima verossimilhança. Já quando λ é muito grande, a variância do modelo é zero, mas o viés é muito elevado e há um *underfitting*. Assim, é necessário que se identifique um valor ótimo de λ que estabeleça um equilíbrio entre a variância e o viés, possibilitando a obtenção de um modelo com boa capacidade preditiva (JAMES et al., 2013). Uma das formas de se escolher o melhor valor de λ é por meio da Validação Cruzada com o método *k-fold.*

No método de Validação Cruzada por *k-fold*, o *conjunto de dados* original é dividido em k partes iguais. O valor de k pode variar entre 1 e n, sendo tipicamente utilizados valores entre 5 ou 10 (KOHAVI, 1995). Uma porção deste *conjunto de dados* (k) é utilizado para o teste e as demais partes são usadas para o treino ($k - 1$). Para cada um dos conjunto*s* de treino, são construídos modelos com diferentes valores de λ. Estes modelos são então testados com o *subsets* de teste e são obtidas as medidas de desempenho, como a área sob a curva ROC (AUC). Este procedimento é repetido k vezes de modo que todos os k grupos tenham funcionado como subgrupos de teste. É feita uma média dos valores para cada λ e é identificado aquele que resulta no modelo com o melhor desempenho, avaliado pela *deviance*, que é uma medida análoga à soma dos quadrados dos resíduos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2013) .

Um aspecto importante do método LASSO é que ele é capaz de zerar os parâmetros do modelo, gerando um modelo esparso. Os valores dos parâmetros não nulos podem ser considerados os mais significativos para a construção do modelo e, por esta razão, o LASSO é reconhecido como uma estratégia interessante para a redução da dimensionalidade de um conjunto de dados por meio da seleção de variáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Uma variação do método LASSO é o *Group Lasso* (MEIER; VAN DE GEER; BÜHLMANN, 2008). Este método foi desenvolvido para problemas nos quais as variáveis preditoras são fatores e é necessário que se selecione grupos de variáveis, não apenas variáveis *dummies* de modo individual.

O estimador do *Group Lasso* é aquele que minimiza:

$$S_\lambda(\boldsymbol{\beta}) = -\left(\sum_{1}^{n} y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) - log\ \left(1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right) - \lambda\sum_{g=1}^{G} s\left(df_g\right)\left|\left|\boldsymbol{\beta}_g\right|\right|2$$

Sendo g os grupos de variáveis e df_g os graus de liberdade do ***g***-ésimo preditor, $s\left(df_g\right)$, é utilizado para garantir que o termo da penalidade é da mesma ordem do número de parâmetros de df_g . Assim como no método LASSO, a escolha do parâmetro λ pode ser realizada por meio de validação cruzada.

3.3.3 Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (*Support Machine Vectors)*, SVM é um método de classificação desenvolvido para problemas de classificação binária. Este método considera que, em um problema de classificação, dado um conjunto de treinamento, existe um hiperplano (equação), não necessariamente linear, que separa perfeitamente as observações de cada uma das classes (CORTES; VAPNIK, 1995). As observações de classes distintas e que estão mais próximas são chamadas de vetores de suporte. Cada um destes vetores estaria a uma distância d do hiperplano e a menor distância entre as observações mais próximas de classes distintas e o hiperplano é denominada margem (M). O hiperplano ótimo seria aquele que maximiza M. Para o problema de hiperplanos não lineares, o SVM utiliza as funções *kernel*, que quantificam as similaridades entre duas observações, para acomodar as separações não lineares entre as duas classes. As funções kernel mais empregadas são as lineares, polinomiais, radiais e sigmoides (HASTIE; TIBSHIRANI; FRIEDMAN, 2009b; JAMES et al., 2013).

Assim, o SVM busca o hiperplano com coeficientes β, de modo que:

$$\underset{\beta_0, \beta_1, \ldots \beta_p,\ \epsilon_1, \ldots.,\ \epsilon_n, M}{maximize} \quad M$$

Sujeito às restrições:

1. $\sum_{j=1}^{p} \beta_j^2 = 1$
2. Para todo: $y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) \geq\ M(1 - \epsilon_i)$, em que: $\epsilon_i\ \geq 0$ e $\sum_{i=1}^{n} \epsilon_j \leq C$

Sendo ϵ_i é um parâmetro que determina onde a *i*-ésima observação está localizada em relação ao hiperplano e em relação à margem. Se $\epsilon_i = 0$, a observação está do lado correto da margem. Se $\epsilon_i > 0$, a observação violou a margem. Se $\epsilon_i > 1$, a observação violou e está

do lado errado do hiperplano. C é um "*tuning parameter*" que corresponde à soma dos valores de ϵ_i e determina o número e a severidade das violações à margem e ao hiperplano que serão aceitas. Quanto maior o valor de C, mais se permite que as observações caiam do lado incorreto da margem.

O hiperplano pode ser escrito como:

$$f(x) = \beta_0 + \sum_{i \epsilon S} \alpha_i \, K\,(x, x_i)$$

Sendo x a nova observação e x_i as observações de treino e K a função *kernel,* por exemplo:

Polinomial: $K\,(x, x_i) = \left(1 + \sum_{j=1}^{p} K\,\left(x_{ij}, x_{ij'}\right)\right)^d$

Radial: $K\,(x, x_i) = exp\left(-\gamma \sum_{j=1}^{p} K\left(x_{ij} - \, x_{ij'}\right)^2\right)$

3.3.4 Métodos Ensemble

Os métodos *Ensemble* são aqueles nos quais modelos simples de predição são combinados com o objetivo de se obter um modelo único mais poderoso. Para isso, são empregados diferentes métodos, como as florestas aleatórias e *boosting*.

3.3.4.1 Florestas Aleatórias

As Florestas Aleatórias (*Random Forests*) são um método *ensamble* no qual o modelo preditivo é construído a partir da combinação das predições de B Árvores de Classificação. O princípio básico das árvores de classificação é a partição recursiva e binária do espaço preditor de treinamento R_j, caracterizada por uma série de variáveis, em espaços menores e mais homogêneos (R_1, ..., R_j), utilizando as variáveis preditoras como pontos de segmentação e regras para classificação dos elementos nos subespaços. O particionamento é repetido nos subespaços até que se alcance um elevado grau de homogeneidade e eles não possam mais ser subdivididos (JAMES et al., 2013). Graficamente, uma árvore de classificação T é entendida, de fato, como uma árvore, na qual o *conjunto de dados* original de treino é a raiz e os subgrupos t são os "nós". Em um primeiro momento, o algoritmo decide qual das p variáveis independentes do *conjunto de dados* terá a primeira divisão binária, ou seja, o melhor ponto que divide o espaço Y da variável resposta. A raiz e os nós são subdivididos de acordo com uma norma *Se-Então*, havendo duas possibilidades de saída (ramos). Os nós

descendentes t são R_1 e R_2 que, quando não podem mais ser divididos, são chamados de nós terminais, ou folhas (classes) (JAMES et al., 2013).

No processo de crescimento de uma árvore de classificação são identificadas, por um processo de otimização, entre as variáveis preditoras $\boldsymbol{x}_i$ e cortes possíveis (s), aquela combinação que conduz, naquele passo, a uma partição $R_1 = \{\boldsymbol{x} \,|\boldsymbol{x}_i < s\}$ e $R_2 = \{\boldsymbol{x} \,|\boldsymbol{x}_i \geq s\}$ com menor "impureza" (maior homogeneidade).

Uma das medidas empregadas para avaliar a "pureza" dos subconjuntos é o Índice de Gini, que é uma medida da variância total entre as K classes (JAMES et al., 2013). Sendo p_{mk} a proporção de observações do grupo treino na $m-ésima$ região que vem da $k-ésima$ classe, o Índice de Gini é dado por:

$$G_m = \sum_{k=1}^{K} p_{mk}\,(1 - p_{mk})$$

Os valores do Índice de Gini variam entre 0 (mais puro) e 0,5 (mais impuro). Assim, valores pequenos de G_m indicam que nos nós predominam observações de uma única classe determinada (JAMES et al., 2013).

O método de Florestas Aleatórias como objetivo diminuir a variância do modelo de árvore de classificação por meio da introdução de algum viés. Para a criação das florestas, diversas árvores de classificação são criadas de modo independente, sendo B o número de árvores criadas. Para isso, por meio da técnica de *bootsrap*, são criadas k amostras *bootstrap*. Para cada amostra k é criada uma árvore de classificação, sendo que, para cada nó, de cada árvore, um grupo m de variáveis preditoras é selecionado por meio de sorteio dentre as p variáveis disponíveis no *conjunto de dados* original. Deste modo, para a divisão em ramos, somente uma das m variáveis sorteadas é utilizada. Este método permite que os preditores de força pequena ou moderada participem da estrutura e que árvores distintas, não correlacionadas, sejam formadas (JAMES et al., 2013).

Para a predição da resposta de uma nova observação, cada árvore "vota" em uma determinada classe, considerando-se o limiar de decisão para as probabilidades, escolhido para classificação, de acordo com sua estrutura, e a floresta escolhe como classe final aquela com o maior número de votos(HASTIE; TIBSHIRANI; FRIEDMAN, 2009b):

$$C_{RF}^{B}(x) = maioria\ dos\ votos\ \{C_b(x)\}_1^B$$

Sendo $C_b(x)$ a classe predita pela árvore *b.*

3.3.4.2 Boosting

Métodos de *Boosting*, assim como as Florestas Aleatórias, são métodos de criação de modelos preditivos que visam a construção de um classificador mais poderoso por meio da agregação de classificadores fracos, como as árvores de decisão ou regressões com poucos parâmetros. Ao contrário do que acontece, entretanto, no método de Florestas Aleatórias, em que as árvores de decisão são formadas de modo independente, no método *Boosting*, as árvores são montadas de modo sequencial, por meio de um aprendizado lento, a partir das árvores anteriores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Um dos mais atuais e populares algoritmos é o *XGBoost (Extreme Gradient Booostting*) (CHEN; GUESTRIN, 2016; FRIEDMAN, 2001) , uma evolução do algoritmo *Gradient Boosting. O XGBoost* é um algoritmo de classificação cujo funcionamento se baseia na combinação de várias árvores de decisão, formando florestas, de modo que, a cada passo, o modelo tenta corrigir o erro do passo anterior.

Considerando-se um determinado conjunto de dados, com n observações e m variáveis

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}\ (|\mathcal{D}| = \ (n_i \boldsymbol{x_i} \in \mathbb{R}^m, y_i \in \mathbb{R})\ ,$$

um modelo ensemble baseado em árvores, com K funções aditivas para predizer a variável resposta pode ser dado por:

$$\hat{y_i} = \emptyset(\boldsymbol{x}_i) = \ \sum_{k=1}^{K} f_k(\boldsymbol{x}_i) \ \text{ sendo que } f_k \ \in \mathcal{F},$$

sendo o espaço das árvores de regressão $\mathcal{F}$, q a estrutura de cada árvore (regras de árvores de decisão), T, o número de folhas, w o peso de cada folha, f_k uma estrutura independente de árvore:

$$\mathcal{F} = \left\{ f(x) = w_{q(x)} \right\} (q: \mathbb{R}^m \rightarrow T, w \ \in \mathbb{R}^T)$$

Neste caso, a função objetivo visa minimizar a função de perda

$$\mathcal{L}(\emptyset) = \sum_{i}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k)\ ,$$

na qual $\boldsymbol{\theta}$ são os parâmetros de otimização, $l(\ )$ é a função de perda (mede a diferença entre o valor estimado e o valore real), $\Omega(\ )$ é a função de regularização (penaliza a complexidade do modelo) e n é o número de observações. Esta função objetivo possui funções como

parâmetros e, por esta razão, pode ser modificada com a introdução dos passos de treinamento:

$$\mathcal{L}^t = \sum_i^n l\left(y_i, \hat{y}_i^{(t-1)}\right) + f_t\left(\boldsymbol{x}_i\right) + \Omega\left(f_t\right)$$

Sendo $\hat{y}_i^{(t)}$ a predição da $i-ésima$ observação na $t-iésima$ interação (t é o número de passos), e f_t funções de aprendizado.

3.3.4.3 Staking

O *Stacking* é um método de construção de modelos preditivos que tem como objetivo a redução de erros de generalização. A ideia central é, a partir de um *conjunto de dados* original, realizar predições com diferentes modelos individuais (*first-level-learners*) a partir de diferentes estratégias de aprendizado de máquina; combinar os valores preditos, os transformando em um novo conjunto de variáveis preditoras; e utilizar as novas variáveis preditoras para ajuste de um novo modelo, utilizando outra abordagem de aprendizado de máquina (second-*level-learner* ou meta-preditor). Este novo modelo, uma vez ajustado, é utilizado para predições (ZHOU, 2012). A fim de evitar *overfitting* do modelo, o treinamento dos *first-level-learners* é feito com validação cruzada e, na segunda etapa, todo o *conjunto de dados* de treino é utilizado para ajuste do meta-preditor.

3.3.5 Treino, Validação e Teste dos Modelos

A validação e teste dos modelos preditivos é uma etapa fundamental, tanto para a avaliação de seu desempenho quanto para a seleção daquele com maior capacidade preditiva. Para isso, diferentes métodos de reamostragem podem ser empregados.

Os métodos de reamostragem envolvem a divisão inicial do conjunto de dados em treino e teste. Idealmente, o conjunto treino é subdividido em novos conjuntos de treino e validação, e o modelo é ajustado. Na etapa seguinte, o modelo ajustado com os dados de treinamento é avaliado utilizando-se o conjunto de teste. Diferentes métodos de reamostragem vêm sendo utilizados para a avaliação do desempenho dos modelos preditivos envolvendo classificação, tais como *Hold out*, Validação Cruzada *k-Fold* e *Bootstrap*. Estes métodos se diferenciam pela forma de partição dos dados nos conjuntos treino e teste (JAMES et al., 2013).

O método *hold-out* é reconhecido como um dos métodos mais simples de reamostragem. Nele, o conjunto de dados original é particionado nos conjuntos treino e validação, sendo o subconjunto de validação chamado de *hold-out*. Tradicionalmente, a divisão dos dados entre os conjuntos de treino e validação é feita considerando-se a proporção de ⅔ destes dados no conjunto de treinamento (KOHAVI, 1995). Os modelos são ajustados no conjunto treino e o desempenho do modelo é avaliado com o conjunto de teste. Uma vez que apenas uma parte dos dados é utilizada no treinamento, o *hold-out* pode gerar uma predição com elevado viés (diferença entre o valor predito e o valor real), principalmente em conjuntos de dados com um número pequeno de observações. Assim, se considera que o emprego deste método conduz a uma estimativa pessimista do erro de predição (JAMES et al., 2013).

A validação cruzada é um dos métodos mais populares para a reamostragem (JAMES et al., 2013). Na validação *k-Fold*, o conjunto de dados é particionado aleatoriamente em k subconjuntos (*folds*) de mesmo tamanho, sendo que um subconjunto *k* é separado para validação. Os conjuntos $k-1$ são utilizados para treino e testadas no grupo teste. Este processo é repetido k vezes, de modo que a cada nova rodada, um conjunto de dados é utilizado para validação. Tipicamente, o número de subconjuntos empregados é de $k=5$ ou $k=10$, sendo que, de acordo com Kohavi (1995), $k=10$ está relacionado a um menor viés (KOHAVI, 1995; XU; GOODACRE, 2018).

Quando o *conjunto de dados* é particionado em um número de subconjuntos correspondente ao número de observações, a validação cruzada é chamada *Leve-one-Out*. Este método, comparado ao *k-Fold*, conduz a uma maior variância, uma vez que o grupo de teste é formado por apenas uma observação (JAMES et al., 2013).

O método *bootstrap*, a partir do *conjunto de dados* inicial, são criados conjuntos de *conjunto de dados*, de mesmo tamanho que o original, por meio de reamostragem com reposição (amostra *bootstrap*). As amostras *bootstrap* formadas por sorteios são utilizadas como conjunto de treino. Os elementos que não foram sorteados para o conjunto de treino, passam a integrar o conjunto de teste. Para cada amostra, o desempenho do modelo é calculado. O processo de sorteio dos subconjuntos, treino e teste é repetido por diversas vezes (JAMES et al., 2013), sendo tradicionalmente empregadas 1.000 repetições (KOHAVI, 1995; XU; GOODACRE, 2018).

3.3.6 Métricas de Desempenho dos Modelos

3.3.6.1 Área sob a Curva

A *Area Under Curve* (AUC), é um método de avaliação de desempenho de modelos preditivos de classificação nos quais a variável resposta é binária estimada a partir da curva ROC (*Receiver Operating Characteristic*). Esta curva se trata de um gráfico de probabilidades que relaciona, a partir da chamada matriz de confusão (Quadro 1), as medidas de sensibilidade e especificidade do modelo (JAMES et al., 2013). Para a construção desta matriz, as classes reais das observações do conjunto de teste, obtidas a partir dos valores de probabilidade estimados e considerando-se um dado ponto de corte para classificação (limiar de decisão ou *threshold*), são comparadas às classes estimadas pelo modelo.

Quadro 1- Matriz de Confusão

Valor Predito	**Valor Verdadeiro**	
	Y=0	Y=1
Y=0	Verdadeiro Negativo (VN)	Falso Negativo (FN)
Y=1	Falso Positivo (FP)	Verdadeiro Positivo (VP)

Considerando-se a matriz de confusão, a sensibilidade é definida como a capacidade do modelo em detectar casos positivos dentre aqueles casos que de fato são positivos. Sendo VP os casos em que os valores positivos classificados corretamente e FN os casos em que os valores positivos foram classificados incorretamente (classificados como negativos), a sensibilidade é dada por:

$$sensibilidade = \frac{VP}{VP + FN}$$

A especificidade, por outro lado, se refere à capacidade do modelo em detectar casos negativos dentre aqueles que de fato são negativos. Sendo VN os casos em que os valores negativos classificados corretamente e FP aqueles em que aos valores negativos foram classificados incorretamente (classificados como positivos), a especificidade é dada por:

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

Para a construção da curva ROC, em um plano cartesiano de vetores unitários, são plotados, no eixo y, os valores verdadeiro-positivos (sensibilidade) e, no eixo x, os valores falso-negativos (1-especificidade) (Figura 2), à medida que o ponto de corte (*threshold*) de classificação varia. Neste plano, a área total é igual a 1 e a AUC se refere ao espaço bidimensional abaixo da curva formada. Os valores de AUC, portanto, podem variar entre 0 e 1 e quanto maiores, melhor o desempenho do modelo. Valores abaixo de 0,5 indicam que o desempenho do modelo foi inferior ao de um classificador aleatório.

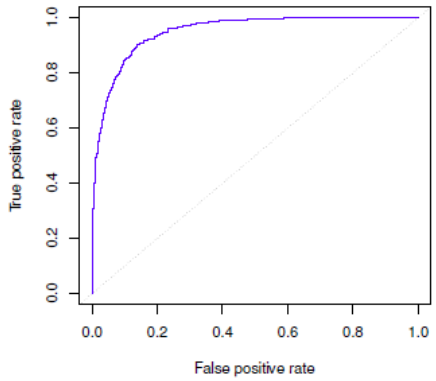

Figura 2- Gráfico da curva ROC
Fonte: James et al. (2013)

3.3.6.2 Acurácia

A acurácia do modelo se trata de uma das mais simples medidas de avaliação de sua capacidade preditiva (JAMES et al., 2013). Ela avalia o número de classificações corretas em relação ao número total de dados. Uma vez que o modelo tenha sido ajustado com as variáveis preditoras, seu desempenho é avaliado por meio de sua aplicação em um conjunto de testes e avaliação do percentual de casos que são corretamente classificados.

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

3.3.6.3 Precisão

A precisão do modelo se refere ao percentual dos valores verdadeiramente positivos em relação ao total de valores classificados como positivos (JAMES et al., 2013).

$$Precisão = \frac{VP}{VP + FP}$$

4 MÉTODOS

O foco deste trabalho é contribuir com o cenário de desenvolvimento de novos imunoterápicos para tratamento do câncer por meio da construção de um modelo preditivo de imunogenicidade de epítopos permita prever, dentre aqueles identificados em amostras tumorais, os que, de fato, são capazes de induzir respostas imunes. Utilizando diferentes estratégia de Aprendizado de Máquina (Florestas Aleatórias, Máquinas de Vetores de Suporte, XGBoost e método *staking*), além de regressões logísticas regularizadas (LASSO e Group LASSO), foram construídos e avaliados diferentes modelos a fim de identificar aqueles com melhor desempenho preditivo. Esta seção descreverá a bases de dados utilizada, as estratégias para treino e validação dos modelos, as métricas para avaliação do desempenho e os procedimentos utilizados para as etapas de ajuste e predição dos modelos. O pré-processamento das variáveis e todas as simulações a seguir detalhadas foram conduzidas no *software R (R Core Team, 2022)*.

4.1 Descrição da Base de Dados

Para o desenvolvimento dos modelos de predição da imunogenicidade de neopítopos tumorais capazes de provocar resposta imune pelas células TCD8+, foi realizada uma busca no banco *Immune Epitope Database* (IEDB) (www.iedb.org), que cataloga dados experimentais dentre outros, de epítopos de células T em humanos no contexto de diferentes patologias.

Uma vez que os neoepítopos tumorais são derivados de proteínas mutadas das células tumorais e expressos pelas moléculas de MHC-I na superfície das células, a busca realizada no IEDB teve como critérios de inclusão: epítopos lineares, formados por sequências de 9 aminoácidos; experimentos de indução da resposta de células T e afinidade com as moléculas de MHC de classe I; células humanas; dados relacionados a epítopos de câncer em geral. Ao final, 1.725 sequências de aminoácidos correspondentes a epítopos tumorais, sendo 506 imunogênicos e 1.219 não imunogênicos, foram utilizados para as etapas de treino e teste dos modelos preditivos.

Para cada um dos epítopos selecionados no IEDB, imunogênicos e não imunogênicos, foram identificados os valores correspondentes a 16 conjuntos de variáveis preditoras,

representativos de propriedades dos peptídeos que poderiam ter influência em seu reconhecimento pelas células T e ativação do sistema imune:

1. **Propensões Físico-químicas:** caracterização de cada um dos 9 aminoácidos da sequência por meio de 531 variáveis que podem estar associadas com a estrutura função da proteína. Estas características foram extraídas da base AAindex (“AAindex: Amino acid index database”, [s.d.]). Neste trabalho, é assumido que peptídeos imunogênicos compartilham similaridades funcionais que são distintas daquelas compartilhadas pelos peptídeos não imunogênicos.

2. **Composição de Aminoácidos:** percentual de cada um dos 20 possíveis aminoácidos encontrados no organismo humano, na sequência de 9 aminoácidos do peptídeo. A composição de aminoácidos poderia indicar a localização do peptídeo dentro da célula tumoral (PARK; KANEHISA, 2003). Esta localização poderia indicar a função do peptídeo, também a sua disponibilidade para ligação às moléculas de MHC I e potencial imunogênico (CASTRO et al., 2022).

3. **Perfil de Pares de Aminoácidos:** percentual de cada um dos 400 possíveis aminoácidos encontrados no organismo humano, em cada um dos 9 aminoácidos da sequência do peptídeo. O perfil de pares de aminoácidos foi considerado relevante porque poderia refletir, por exemplo, o papel das interações entre as propriedades individuais de cada aminoácido em uma sequência (CHEN et al., 2007; KRINGELUM et al., 2013). Além disso, há evidências de que a composição de pares de aminoácidos resultaria em alguma melhoria da predição da localização subcelular dos peptídeos (PARK; KANEHISA, 2003) e também de imunogenicidade (CAI et al., 2003).

4. **Perfil Esparso:** representação de cada um dos aminoácidos da sequência por uma *string* binária, considerando os 20 possíveis aminoácidos presentes no organismo humano. Um bit é codificado pelo número 1 e os outros por 0. Esta variável está, portanto, relacionada à composição de aminoácidos.

5. **Escore Blossum62:** escore representativo da probabilidade de substituição, de cada um dos 9 aminoácidos da sequência por cada um dos 20 aminoácidos possíveis encontrados no corpo humano(HENIKOFF; HENIKOFF, 1992). Esta variável carrega informações evolucionárias que seriam relevantes para a predição de imunogenicidade de neoepítopos uma vez que as mutações nas proteínas das células

tumorais é que permitem o seu reconhecimento como um elemento não-próprio do organismo.

6. **Energia de Ligação entre Pares de Aminoácidos:** para cada um dos 9 aminoácidos são identificados os valores de energia de ligação entre os aminoácidos do peptídeo e os resíduos de um modelo estrutural de MHC. Para cada um dos aminoácidos são apresentados 40 valores de energia de ligação (SAETHANG et al., 2013). Esta variável permitiria, portanto, distinguir entre os epítopos que se melhor se ligam ao MHC ou ao TCR (imunogênicos) e aqueles que não tem uma boa ligação (não imunogênicos) (SCHUELER-FURMAN et al., 2000).

7. **Similaridade Molecular Topológica Quântica:** para cada um dos 9 aminoácidos da sequência, são indicados os escores das componentes principais para 4 métodos distintos de criação de índices (CBFQ, CDFQ, CUFQ, ADFQ) que caracterizam a molécula a partir da topologia da densidade eletrônica (HEMMATEENEJAD; YOUSEFINEJAD; MEHDIPOUR, 2011). A similaridade molecular topológica quântica é uma variável que permitiria classificar, a partir das propriedades eletrônicas dos resíduos, os epítopos com melhor e pior estabilidade de ligação com o receptor das células TCD8+.

8. **Perfil Global de Composição, Transição e Distribuição:** para cada um dos 9 aminoácidos da sequência são identificados os valores relacionados a 3 subcategorias de 13 propriedades físico-químicas e estruturais[8], considerando-se a composição, frequência e distribuição da subcategoria de propriedade na cadeia (CUI et al., 2007). Esta variável foi selecionada porque poderia ser representativa da conformação 3D de uma proteína e de sua função (DUBCHAK et al., 1995; GOVINDAN; NAIR, 2011). Estes aspectos seriam relevantes para predizer o seu acoplamento à cavidade das moléculas de MHC I e também ao receptor do linfócito TCD8+.

9. **Autocorrelação de Moran:** para cada peptídeo, é indicada, para 8 diferentes propriedades [2] a diferença entre o valor daquela propriedade naquele ponto da cadeia e o valor médio da propriedade para a cadeia (CHEN et al., 2020). São considerados 7 pontos no comprimento da cadeia do peptídeo. Estudos indicam que esta variável

[8] Propriedades físico-químicas e estruturais: hidrofobicidade (PRAM900101; ARGP820101; ZIMJ680101; PONP930101; CASG920101; ENGD860101; FASG890101), polaridade, estrutura secundária, carga; polarizabilidade; acessibilidade a solventes; volume normalizado de Van der Walls.

poderia predizer a estrutura e a função de uma proteína (ONG et al., 2007) e, portanto, foi selecionada para este estudo, em que se se considera que tanto a estrutura quanto a função dos epítopos seja diferente entre os grupos imunogênicos em não imunogênicos.

10. **Autocorrelação de Geary:** para cada peptídeo, é indicada para 8 diferentes propriedades,[2] o quadrado da diferença entre o valor daquela propriedade naquele ponto da cadeia e o valor médio da propriedade para a cadeia (CHEN et al., 2020). São considerados 7 pontos no comprimento da cadeia do peptídeo. Estudos indicam que esta variável poderia predizer a estrutura e a função de uma proteína (ONG et al., 2007) e, portanto, foi selecionada para este estudo, em que se se considera que tanto a estrutura quanto a função dos epítopos seja diferente entre os grupos imunogênicos em não imunogênicos.

11. **Autocorrelação de Moreau-Broto:** para cada peptídeo, é indicada a distribuição dos valores de 8 propriedades[9] ao longo da cadeia. São considerados 7 pontos no comprimento da cadeia do peptídeo (CHEN et al., 2020). Esta variável descreve a relação global entre os resíduos (AL-BARAKATI et al., [s.d.]). Estudos indicam que esta variável poderia predizer a estrutura e a função de uma proteína (ONG et al., 2007) e, portanto, foi selecionada para este estudo, em que se se considera que tanto a estrutura quanto a função dos epítopos seja diferente entre os grupos imunogênicos em não imunogênicos e, portanto, se constitua como uma variável preditora adequada.

12. **Quasi-sequence:** variável composta por duas variáveis: matriz de distâncias físico-químicas de *Schneider* e matriz de distâncias químicas de *Granthan*. Cada uma delas é composta por um vetor com 28 valores, sendo que os 20 primeiros se referem ao efeito da composição de aminoácidos e os 8 últimos referem à distribuição de uma propriedade ao longo de 8 posições na cadeia do peptídeo. Uma vez que esta variável é capaz de predizer a localização subcelular dos peptídeos e a sua função (CHOU, 2000), espera-se que ela possa ser capaz de diferenciar os grupos de epítopos imunogênicos e não imunogênicos.

13. **Composição de Pseudo-aminoácidos:** o peptídeo é caracterizado por um vetor com 28 elementos. Os 20 primeiros elementos se referem à frequência de cada um dos 20

[9] (hidrofobicidade, index de flexibilidade, polarizabilidade, energia livre, área de superfície livre, volume do resíduo, parâmetros estéricos e mutabilidade relativa).

possíveis aminoácidos na sequência do epítopo, e os 8 restantes se referem a fatores correlacionados que refletem efeitos incrementais nas propriedades hidrofobicidade, hidrofilicidade e a massa da cadeia lateral, decorrentes da ordem dos aminoácidos na sequência (CHOU, 2001). Estudos anteriores indicam que a composição de pseudo-aminoácidos pode prever a classe estrutural das proteínas (CHEN et al., 2006). Deste modo, se trata de uma variável que pode contribuir para a predição do acoplamento entre os neoepítopos e a cavidade da molécula de MHC e também entre os neoepítopos e os receptores dos linfócitos TCD8+.

14. **Composição de Aminoácido pseudo-anfifílicos:** o peptídeo é caracterizado por um vetor com 24 elementos. Os 20 primeiros elementos se referem à frequência de cada um dos 20 possíveis aminoácidos na sequência do epítopo, e os 4 restantes se referem a valores correspondentes à distribuição de aminoácidos hidrofóbicos e hidrofílicos ao longo da cadeia (CHOU, 2005). Esta distribuição de aminoácidos hidrofóbicos e hidrofílicos ao longo da sequência impacta a estrutura da proteína. Assim, considerando-se que os peptídeos imunogênicos podem se diferenciar dos não imunogênicos por seu acoplamento estrutural às moléculas de MHC I e também aos receptores de células T, esta variável foi incorporada ao estudo. Outro ponto importante é que a hidrofobicidade é considerado como uma das principais características que determinam a ligação entre uma proteína e seus ligantes (LIMONGELLI; MARINI; BELLAZZI, 2015).

15. **Acessibilidade relativa a solventes:** para cada um dos 9 aminoácidos da sequência, é indicado um valor de acessibilidade a solventes daquele aminoácido, normalizado pelo valor máximo disponível para aquele resíduo (ADAMCZAK; POROLLO; MELLER, 2005). Esta variável vem sendo empregada tanto para a predição de estruturas secundárias de proteínas quanto para a predição de mutações em sua estrutura (TIEN et al., 2013).

16. **Predição de estrutura secundária:** para cada um dos 9 aminoácidos é apresentado um vetor binarizado, com 3 elementos que representa o tipo de estrutura secundária do peptídeo: α-hélice (1,0,0), folha-β (0,1,0) e sequências sem estrutura (0,0,1) (ADAMCZAK; POROLLO; MELLER, 2005).

A Tabela 1 apresenta uma síntese das 16 variáveis que foram empregadas para o desenvolvimento do modelo preditivo de imunogenicidade de neoepítopos.

Tabela 1- Detalhamento dos grandes conjuntos de variáveis potencialmente preditoras de imunogenicidade de epítopos tumorais.

Identificador	Grupo de Variáveis	Número total de variáveis isoladas	Dimensão
D1	Propensões Físico-químicas	4780	531 x 9
D2	Composição de Aminoácidos	20	1 x 20
D3	Perfil de Pares de Aminoácidos	400	20 x 20
D4	Perfil Esparso	180	20 x 9
D5	Escore Blossum62	180	9 x 20
D6	Energia de Ligação entre Pares de Aminoácidos	360	40 x 9
D7	Similaridade Molecular Topológica Quântica	189	4 (7 x9; 6x 9; 3x9; 2x9)
D8	Perfil Global de Composição, Transição e Distribuição	273	13 x 21
D9	Autocorrelação de Moreau-Broto	56	8 x 7
D10	Autocorrelação de Geary	56	8 x 7
D11	Autocorrelação de Moran	56	8 x 7
D12	Ordem de quasi-sequence	56	2 (20 + 8; 20 + 8)
D13	Composição de Pseudo-aminoácidos	28	2 (20 + 8; 20 + 8)
D14	Composição de Aminoácido pseudo-anfifílicos	24	(20 + 4)
D15	Acessibilidade relativa a solventes	9	9
D16	Predição de estrutura secundária	27	9 x 3

4.2 Avaliação do Desempenho dos Modelos

Para a avaliação do desempenho e seleção dos modelos gerados, foram empregadas como métricas, os valores da Área sob a Curva (AUC) e sensibilidade (descritos na subseção 3.3.5). A AUC foi selecionada por ser o método mais empregado para avaliar o desempenho geral de um modelo preditivo de classificação (JAMES et al., 2013) e, uma vez que o desenvolvimento de imunoterápicos envolve a identificação de epítopos que são capazes de induzir a resposta imune, a sensibilidade também foi empregada para a seleção dos modelos.

As estratégias de reamostragem *hold-out*, validação cruzada *k-fold* e *bootsrap* foram empregadas para treino e validação dos modelos, conforme descrito a seguir. Para a execução do *hold-out*, o *conjunto de dados* original foi particionado, aleatoriamente, nos grupos treino e teste. O particionamento considerou a proporção de 80% das observações no grupo treino e 20% das observações no grupo teste. Os modelos foram ajustados com os dados do grupo treino e o desempenho foi avaliado, aplicando o modelo às observações do grupo teste.

Para execução das validações cruzadas *k-fold*, o *conjunto de dados* original foi dividido, aleatoriamente, em 10 subconjuntos ($k = 10$), utilizado a função $cross\_kfold$, do pacote $modelr$, do *software* R. A divisão dos subconjuntos considerando $k = 10$ foi selecionada por ser este o valor mais frequentemente utilizado em trabalhos anteriores e por ter sido descrito como aquele capaz de conduzir a modelos com melhor desempenho (BORRA; DI CIACCIO, 2010; KOHAVI, 1995). Um dos conjuntos k foi separado para o teste e o restante $k - 1$, para treino. Este processo foi repetido 10 vezes, com novas seleções de conjuntos de treino e teste. A cada repetição, o desempenho do modelo foi estimado pelo valor de AUC. O desempenho do modelo foi dado pela mediana dos valores de AUC individuais de cada rodada, sendo que o melhor modelo individual foi selecionado para as etapas seguintes.

Outra estratégia de reamostragem utilizada para avaliação do desempenho dos modelos foi o *bootsrap*, um método considerado poderoso e associado a uma menor variância. Para a sua execução, foi empregada a função $bootstrap$, do pacote $modelr$, com 1.000 amostras de *bootstrap* $(k = 1000)$ sorteadas, com repetição, a partir do *conjunto de dados* original. Cada uma destas amostras *boostrap* apresentava o mesmo número de observações do conjunto original e foi particionada, por meio o método *hold out*, em conjuntos de treino e teste, na proporção 80/20. Para cada uma das amostras *boostrap* foi, portanto, ajustado um modelo

preditivo e calculada a AUC. O desempenho geral do modelo *boostrap* foi dado pela mediana da AUC dos conjuntos individuais.

4.3 Seleção de variáveis

Para a construção de um modelo capaz de predizer se um determinado peptídeo é imunogênico ou não imunogênico, foram selecionadas 5.913 variáveis independentes, agrupadas em 16 conjuntos, que poderiam ter alguma contribuição no processo de ligação dos peptídeos à molécula de MHC I e ao seu reconhecimento pelos receptores de células T. Este número de variáveis é superior ao número de observações ($n = 1.725$), caracterizando um *conjunto de dados* de alta dimensionalidade. Modelos com um elevado número de variáveis, em geral, possuem um excelente ajuste ao conjunto de dados de treino. Por outro lado, apresentam uma elevada variância e, portanto, uma capacidade reduzida de generalização para outro conjunto de dados, como aquele do conjunto de validação ou teste (HASTIE; TIBSHIRANI; FRIEDMAN, 2009b; JAMES et al., 2013).

No contexto deste trabalho, em que a variável reposta Y assume dois valores, 0 para peptídeos não imunogênicos e 1 para peptídeos imunogênicos, foram ajustados modelos logísticos, conforme descrição apresentada no subitem 3.3.1. Para identificação das variáveis mais importantes para a discriminação dos peptídeos nos grupos imunogênicos e não imunogênicos, a fim reduzir a variância do modelo final, a estes modelos logísticos foram associados dois tipos de penalização: LASSO e *Group LASSO*.

4.3.1 Modelos Logísticos com Penalização LASSO

Este método, descrito na subseção 3.3.2, vem sendo utilizado em diferentes estudos envolvendo bancos de dados de alta dimensionalidade relacionados ao câncer (O'SHEA et al., 2021; TIAN; CHEN; JIANG, 2023; YU et al., 2021) e foi selecionado por ser capaz de garantir que as variáveis menos importantes tenham seu coeficiente zerado e também por resolver o problema da multicolinearidade, típico dos modelos com muitas variáveis preditoras.

O ajuste do modelo logístico com penalização Lasso foi realizado com o pacote $glmet$ (FRIEDMAN; HASTIE; TIBSHIRANI, 2010) e a função $\mathrm{cv.glmet}\,(\,)$. Esta função permite que o ajuste do modelo seja realizado a partir da seleção do melhor parâmetro de penalização λ por meio de validação cruzada.

Para as simulações, foram utilizados: parâmetro $\alpha = 1$, que indica a execução de uma regularização lasso, $\mathrm{nfolds} = 10$ e 100 modelos. Os modelos selecionados foram aqueles construídos com o λ que conduziu ao menor erro de predição. Estes modelos foram empregados na predição com o conjunto teste, utilizando a função $\mathrm{predict}()$.

4.3.2 Modelos Logísticos com Penalização *Group LASSO*

Outro método empregado de penalização avaliado foi o *Group LASSO* (YANG; ZOU, 2015), uma variação do LASSO. Este método é adequado às análises em que alguns preditores fazem parte de grupos pré-definidos e a seleção das variáveis mais importantes é feita considerando tais grupos e não mais as variáveis de modo individual (HASTIE; TIBSHIRANI; FRIEDMAN, 2009b; JAMES et al., 2013). Em função das características do *conjunto de dados* utilizado neste trabalho, em que algumas características do epítopo são interpretadas somente quando tomadas em grupo (D1,D3, D4, D5, D6, D7, D8), o *Group Lasso* foi também avaliado como estratégia para a seleção das variáveis preditivas mais importantes.

O ajuste do modelo logístico com penalização *Group* LASSO foi realizado com o pacote $gglasso$ (YANG; ZOU, 2015). Primeiramente, as variáveis contínuas foram normalizadas com a função $\mathrm{scale}()$ e, em seguida, as variáveis preditivas foram codificas de acordo com os grupos pré-existentes. Para todas as variáveis de um grupo, foi atribuído um mesmo índice numérico. Para o ajuste do modelo, foi utilizada a função $\mathrm{cv.glmet}()$, e para a predição, a função $\mathrm{scale}()$, seguindo os mesmos procedimentos descritos anteriormente para a penalização LASSO.

Para o treino e validação dos modelos gerados com os métodos LASSO e Group LASSO, foram empregadas as estratégias *hold-out*, validação cruzada e *bootstrap*, descritas na seção 4.2.

A fim de identificar o efeito da redundância de informações, além das simulações realizadas com toda a base de dados, foram conduzidas simulações combinando os diferentes grupos de variáveis (D1 a D16) de modo incremental. Para isso, iniciou-se com a simulação de um modelo apenas com a variável D1, seguindo-se de outra simulação com as variáveis D1 e D2 e assim, sucessivamente. O Quadro 2 apresenta as combinações de variáveis que foram utilizadas para a construção dos modelos logísticos com penalização LASSO e *Group* LASSO.

Quadro 2- Combinações diretas de variáveis preditora

D1
D1+D2
D1+D2+D3
D1+D2+D3+D4
D1+D2+D3+D4+D5
D1+D2+D3+D4+D5+D6
D1+D2+D3+D4+D5+D6+D7
D1+D2+D3+D4+D5+D6+D7+D8
D1+D2+D3+D4+D5+D6+D7+D8+D9
D1+D2+D3+D4+D5+D6+D7+D8+D10
D1+D2+D3+D4+D5+D6+D7+D8+D10+D11
D1+D2+D3+D4+D5+D6+D7+D8+D10+D11+D12
D1+D2+D3+D4+D5+D6+D7+D8+D10+D11+D12+D13
D1+D2+D3+D4+D5+D6+D7+D8+D10+D11+D12+D13+D14
D1+D2+D3+D4+D5+D6+D7+D8+D10+D11+D12+D13+D14+D15
D1+D2+D3+D4+D5+D6+D7+D8+D10+D11+D12+D13+D14+D15+D16

Os modelos construídos com cada um destes métodos e com diferentes estratégias de validação tiveram seu desempenho comparado e as variáveis do modelo que apresentou melhor desempenho foram empregadas para o desenvolvimento do modelo preditivo final utilizando diferentes métodos de Aprendizado de Máquina, conforme descrito a seguir.

4.4 Modelos de Aprendizado de Máquinas

4.4.1 Florestas Aleatórias

Conforme descrito na subseção 3.3.4.1, as Florestas Aleatórias (RF) são um método de aprendizado de máquina do tipo *bagging* em que um modelo preditivo classificador de elevado desempenho é construído a partir de árvores de classificação. Este método vem sendo utilizado em diferentes estudos preditivos relacionados ao câncer (TOTH et al., 2019; TSENG et al., 2019) e, em função de sua potencialidade para a geração de modelos com menor variância, foi selecionado para este trabalho.

Para a criação dos modelos preditivos com o método de Floresta Aleatória, foi utilizada a função $randomForest(\ )$ do pacote $ranger$. Utilizou-se, como parâmetros, 500 árvores

($ntree = 500$) um número de variáveis sorteadas em cada nó igual a $\sqrt{p}$ ($mty = \sqrt{p}$) , sendo p o número de parâmetros. Dentre os diferentes modelos gerados, aquele com a maior AUC foi utilizado para a predição utilizando a função $predict\ (\ )$ e os dados do conjunto de teste. A validação do modelo foi realizada com o método *boostrap*, descrito no item 4.2. A importância de cada variável para o modelo preditivo foi obtida e apresentada com auxílio das funções $importance\ (\ )$ e $varImpPlot\ (\ )$.

4.4.2 Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (SVM) são um método de classificação cujo objetivo é identificar um hiperplano ótimo que possa maximizar a distância entre as observações de duas classes em um espaço multidimensional de variáveis preditoras. Assim, o SVM pode ser utilizado em casos em que as observações de cada uma das classes não estejam perfeitamente separadas por um hiperplano linear.

Uma vez que os neoepítopos são gerados por mutações no genoma e que, de um modo geral, tais mutações correspondem a pequenas mudanças na sequência de aminoácidos dos peptídeos formados, espera-se que a variabilidade entre os grupos imunogênico e não imunogênico não seja modelada por um hiperplano linear. Por este motivo, o método de SVM foi selecionado para compor o presente estudo. Além disso, as SVMs vem sendo empregadas em diferentes trabalhos relacionados à predição de epítopos (SAETHANG et al., 2013; TUNG et al., 2011).

Para a construção do modelo foi utilizado o pacote e_1071 e a função $tune(\ )$ com o método svm. Esta função permite que os hiperparâmetros sejam ajustados utilizando um intervalo de valores. Com os dados do grupo de teste, foram avaliados diferentes valores de custo ($cost = 1 * 10\text{^}(-2{:}\,2)$), função Kernel ($kernel = c("radial","polynomial","sigmoid")$). Dentre os diferentes modelos gerados, aquele com a maior AUC foi utilizado para a predição utilizando a função $predict\ (\ )$ e os dados do conjunto de teste. Para validação do modelo, foi utilizado o método *boostrap*, descrito na seção 4.2

4.4.3 XGBoost

O XGBoost é algoritmo de aprendizado de máquina baseado na estratégia de *boosting*, conforme descrito na seção 3.3.4.2. Este método foi selecionado para compor o presente

estudo em função da característica de aprendizado sequencial, que poderia evitar o *overfitting* do modelo e, portanto, conduzir a uma melhor capacidade preditiva (HASTIE; TIBSHIRANI; FRIEDMAN, 2009b). Além disso, porque o método, quando comparado a outros classificadores, como Máquinas de Vetores de Suporte, Florestas Aleatórias, *K*-Vizinhos mais próximos, Regressão Logística e Árvore de Decisão, apresentou melhor desempenho em estudos relacionados em estudos de classificação de tumores (LI et al., 2022) e priorização de neoantígenos (ZHOU et al., 2019).

Para ajuste do modelo, foram utilizados o pacote $xgboost$ e a função $xgboost\ (\ )$. Como parâmetros foram utilizados um número de árvores igual a 1.000 $(nrounds = 1000)$ e profundidade máxima de cada árvore igual a 6 $(max.depth = 6)$. Dentre os diferentes modelos gerados, aquele com a maior AUC foi utilizado para a predição utilizando a função $predict\ (\ )$. A validação do modelo final foi realizada com o método *boostrap*, conforme apresentado na seção 4.2.

4.4.4 Stacking Model

Conforme apresentado na seção 3.3.4.3, a combinação de diferentes modelos de aprendizado de máquina em um meta-preditor pode ser uma estratégia interessante para geração de um modelo preditivo com melhor desempenho quando comparado aos modelos individuais. Assim, a fim de verificar se estratégia conduziria a um preditor mais robusto de imunogenicidade, foi conduzido um processo de stacking em duas etapas.

Na primeira etapa, a partir do *conjunto de dados* original, foram construídos modelos de predição foram construídos modelos de predição de imunogenicidade com os métodos XGBoost e SVM e utilizando validação cruzada ($k = 10$) para treino e teste, conforme descrito na seção 4.2. Os parâmetros para construção destes modelos foram os mesmos descritos nas seções anteriores. Para cada um dos métodos de aprendizado de máquina, os valores preditos para a variável resposta, para cada um dos modelos gerados na validação cruzada, foram “empilhados” e passaram a representar uma variável. As variáveis compostas pelas predições obtidas com os métodos XGBoost e SVM foram reunidas em uma matriz conjuntamente com a variável resposta do *conjunto de dados* original. Esta matriz foi tomada como um novo *conjunto de dados* e o método RF foi empregado como *second-level-learner* e, para validação, foi utilizado o método de reamostragem *boostrap*.

5 RESULTADOS E DISCUSSÃO

5.1 Seleção de Variáveis

Um dos grandes desafios no contexto do desenvolvimento de imunoterápicos contra o câncer é identificar, dentre os vários fragmentos de proteínas apresentados pelas células tumorais, aqueles que são imunogênicos, ou seja, que são capazes de desencadear uma resposta imune.

A ativação das células de defesa T CD8+ por um peptídeo tumoral envolve a ligação do peptídeo à molécula de MHC, apresentação do complexo MHC-peptídeo na superfície celular e ligação deste complexo com os receptores das células citotóxicas. Estes processos de ligação estão relacionados tanto às propriedades físico-químicas quanto estruturais dos peptídeos, que podem ser consideradas como potenciais preditoras da imunogenicidade. Verifica-se, entretanto, um elevado número de possíveis características a serem incorporadas no modelo preditor, o que poderia conduzir à elevada complexidade e *overfitting*. Além disso, muitas variáveis tornam o modelo excessivamente complexo e dificulta a sua aplicação prática, dados os custos operacionais de se obter os valores de todas as variáveis para novas observações.

Deste modo, a fim de construir um modelo preditor de imunogenicidade de epítopos, na primeira etapa deste estudo, o método de regressão logística com as penalizações LASSO e *Group*-LASSO foram empregados a fim de obter modelos esparsos apenas com variáveis cujos coeficientes não foram zerados no processo de maximização da função de verossimilhança. A Tabela 2 apresenta os valores da mediana das métricas de desempenho para os modelos logísticos construídos com diferentes combinações de variáveis para o método LASSO após o processo de treino e validação utilizando $\lambda = 0.03384597$ e a estratégia de *bootstrap*.

Tabela 2- Valores das medianas das principais métricas de desempenho dos modelos gerados com o método LASSO a partir de diferentes combinações de grupos de variáveis preditoras e 1.000 corridas *bootstrap*

Combinações de grupos de variáveis preditoras	AUC	Especificidade	Sensibilidade	Variáveis selecionadas (n)
D1	0,6328	0,5962	0,6748	37
D1-D2	0,6420	0,6078	0,6800	49
D1-D3	0,6631	0,6264	0,7044	59
D1-D4	0,6543	0,6122	0,6981	44
D1-D5	0,6547	0,6108	0,7010	44
D1-D6	0,6489	0,6082	0,6958	36
D1-D7	0,6473	0,5943	0,7009	29
D1-D8	0,6397	0,5842	0,6981	26
D1-D9	0,6525	0,6085	0,6971	34
D1-D10	0,6529	0,6085	0,6985	34
D1-D11	0,6480	0,6000	0,7010	26
D1-D12	0,6566	0,6075	0,7077	33
D1-D13	0,6497	0,6020	0,7064	30
D1-D14	0,6516	0,6070	0,7030	30
D1-D15	0,6508	0,6019	0,6990	30
D1-D16	0,6512	0,6022	0,7077	30

A Tabela 3 apresenta a descrição da distribuição dos valores de AUC, Especificidade, Sensibilidade e Número de Variáveis Selecionadas, considerando-se média, desvio padrão, mediana, valores mínimo e máximo obtidos para os modelos obtidos com as diferentes combinações de *conjunto de dados* e a regressão LASSO.

Tabela 3-Sumário das principais métricas de desempenho do grupo de modelos gerados com as diferentes combinações de variáveis e o método LASSO

	AUC	Especificidade	Sensibilidade	Variáveis selecionadas (n)
Média	0,6498	0,6049	0,6983	35,6875
Desvio padrão	0,0071	0,0092	0,0090	9,0754
Mediana	0,6510	0,6073	0,7000	33,5000
Mínimo	0,0047	0,0074	0,0044	5,1891
Máximo	0,6328	0,5842	0,6748	26,0000

A análise das Tabelas 2 e 3 permite verificar que, assim como observado no trabalho de Zhang et al. (2015) Zhang et al. (2012) (ZHANG et al., 2012, 2015), o incremento sequencial dos grandes conjuntos de variáveis ao modelo de regressão resultou em modelos preditivos de imunogenicidade com pouca diferença entre seus desempenhos. Isso pode estar relacionado ao fato de que, nem todas as variáveis incialmente propostas como

potencialmente preditoras da classificação dos epítopos como imunogênicos ou não imunogênicos, são de fato, aquelas com efeito significativo, ou também porque muitas das variáveis são correlacionadas entre si.

O valor mínimo de AUC observado foi de 0,6328 para o conjunto composto apenas pelas variáveis associadas à D1 (Propensões físico-químicas) e o valor máximo foi de 0,6631, para o conjunto composto pelas variáveis associadas a D1 (Propensões físico-químicas, D2 (Composição de Aminoácidos) e D3 (Perfil de Pares de Aminoácidos). Neste caso, o valor de sensibilidade, ou seja, a capacidade do modelo em detectar casos positivos dentre aqueles casos que de fato são positivos foi de 0,6748.

Em relação ao número de variáveis preditoras, o modelo com melhor desempenho (D1+D2+D3), com AUC= 0,6631, foi construído com 59 variáveis, sendo que 30 delas faziam parte do grupo D1. É interessante observar que este modelo privilegiou as propensões físico-químicas dos aminoácidos individuais, a composição de aminoácidos e o perfil de pares de aminoácidos. Este resultado indica que, de fato, os peptídeos imunogênicos e não imunogênicos apresentam entre si diferenças nas propriedades físico-químicas que são capazes de diferenciar os dois grupos. A importância das variáveis associadas às propensões físico-químicas para a predição de imunogenicidade de epítopos também foi observada por ZHANG et al. (2015), que identificou que apenas este conjunto de variáveis conduziu a um modelo com AUC igual a 0,738.

É interessante também observar que as variáveis selecionadas do grupo D1, em sua maioria, correspondiam a propriedades dos aminoácidos das posições 2, 8 e 9 na cadeia do peptídeo[10]. Estas posições também foram identificadas nos trabalhos de Tung et al. (TUNG et al., 2011) e estão nas extremidades das cadeias dos peptídeos, sendo locais de ancoragem do epítopo à molécula de MHC I.

A Tabela 4 apresenta os valores das quatro variáveis que mais contribuem positivamente para a classificação do epítopo como imunogênico e, na Tabela 5 aquelas que mais contribuem negativamente com esta classificação.

[10] Neste estudo, foram consideradas peptídeos formados por uma cadeia de nove aminoácidos.

Tabela 4- Variáveis que mais contribuem para a classificação dos epítopos como imunogênicos

Identificador	Variáveis	*Odds Ratio*
D3	VT	1,588153131
	GI	1,268444866
	VC	1,130984552
D1	AA1. AURR9080101	1,146756347

Tabela 5- Variáveis que mais contribuem positivamente para a classificação dos epítopos como não imunogênicos

Identificador	Variáveis	*Odds Ratio*
D2	S	0,47154536
D3	TG	0,622576894
D3	MK	0,632965035
D1	AA6. PRAM820103	0,65138186

Verifica-se três das quatro variáveis que mais contribuíram individualmente e positivamente para aumento da chance de o peptídeo ser classificado como imunogênico se tratam de pares de aminoácidos (VT, GI e VC). A Valina, presentes em dois destes 3 pares de aminoácidos, bem como a Isoleucina já foram relatadas como relevantes para o processo de ancoragem do peptídeo à molécula de MHC I, o que é uma etapa relevante para a sua apresentação aos linfócitos TCD8+.

O outro método de penalização avaliado neste trabalho foi a penalização *Group*-Lasso. Para as combinações diretas das categorias D1 a D16, foram considerados 919 grupos de variáveis. A Tabela 6 apresenta os valores da mediana das métricas de desempenho dos modelos logísticos construídos com o método *Group*-LASSO após o processo de treino e validação utilizando $\lambda = 0{,}009392356$ e a estratégia de *bootstrap*.

Tabela 6- Valores das medianas das principais métricas de desempenho dos modelos gerados com o método LASSO a partir de diferentes combinações de grupos de variáveis preditoras e 1 000 corridas bootstap

Combinações de grupos de variáveis preditoras	AUC	Especificidade	Sensibilidade	Número de variáveis preditoras selecionadas
D1	0,6451	0,6373	0,6571	216
D1-D2	0,6511	0,6422	0,6574	258
D1-D3	0,6506	0,6394	0,6595	258
D1-D4	0,6515	0,6421	0,6636	258
D1-D5	0,6491	0,6460	0,6599	307
D1-D6	0,6519	0,6381	0,6667	325
D1-D7	0,6293	0,6162	0,6459	117
D1-D8	0,6443	0,6359	0,6571	236
D1-D9	0,6884	0,6811	0,6989	442
D1-D10	0,6958	0,6893	0,7045	490
D1-D11	0,6874	0,6792	0,6944	390
D1-D12	0,6881	0,6804	0,6975	334
D1-D13	0,6874	0,6771	0,6979	334
D1-D14	0,6881	0,6820	0,6976	326
D1-D15	0,6867	0,6774	0,6954	317
D1-D16	0,6881	0,6820	0,6971	317

Na Tabela 7 é apresentada a descrição da distribuição dos valores de AUC, Especificidade, Sensibilidade e Número de Variáveis Selecionadas, considerando-se média, desvio padrão, mediana, valores mínimo e máximo, obtidos para os modelos obtidos com as diferentes combinações de *conjunto de dados* e a regressão *Group*-LASSO.

Tabela 7-Sumário das principais métricas de desempenho do grupo de modelos gerados com as diferentes combinações de variáveis e o método *Group*-LASSO

	AUC	Especificidade	Sensibilidade	Número de Variáveis Selecionadas
Média	0,6677	0,6591	0,6782	307,8125
Desvio padrão	0,0225	0,0237	0,0209	88,5289
Mediana	0,6693	0,6616	0,6806	317,0000
Mínimo	0,6293	0,6162	0,6459	117,0000
Máximo	0,6958	0,6893	0,7045	490,0000

A análise das Tabelas 6 e 7 permite verificar que o menor valor de AUC foi 0,6293, observado para o conjunto composto pelo menor número de variáveis (117 variáveis), com a combinação D1 a D7. Já o maior valor de AUC foi 0,6958, obtido para o conjunto de 490 variáveis, resultante da combinação dos grupos D1 a D10 (85 grupos). Este conjunto também foi aquele para o qual se observou maior valor de sensibilidade, 70,45%.

Das 531 propensões físico-químicas originais, 30 fizeram parte do melhor modelo. É interessante observar que também se mostraram importantes para a composição o modelo: composição de aminoácidos; matriz blosum62; potencial de contato entre pares de aminoácidos; similaridade molecular topológica quântica; transição das propriedades hidrofobicidade (PONP930101 e FASG890101); volume de Wan der walls; polaridade e estrutura secundária e distribuição da hidrofobicidade (PONP930101 e FASG890101); autocorrelação de Moran (índice de flexibilidade, parâmetro de polarizabilidade, volume do resíduo e mutabilidade relativa) e autocorrelação de Geary (média das escalas de hidrofobicidade, índice de flexibilidade, parâmetro de polarizabilidade e mutabilidade relativa). Assim, o modelo ajustado com o *Group*-Lasso, ao contrário do observado para o ajuste com o modelo LASSO, identificou que as variáveis relacionadas à estrutura do peptídeo e também à matriz evolucionária foram relevantes para a predição de sua imunogenicidade. Este aspecto é importante, dado que a conformação da cadeia de aminoácidos que forma o epítopo é que permite seu encaixe físico na fenda da molécula de MHC e também no receptor das células TCD8+.

A análise dos 490 coeficientes das variáveis que compuseram o modelo permite verificar que 250 deles são positivos, ou seja, contribuem positivamente com a classificação dos peptídeos como imunogênicos, e 239 são negativos, contribuindo com a classificação dos peptídeos como não imunogênicos. A Tabela 8 apresenta os valores das 04 variáveis que mais contribuem positivamente para a classificação do epítopo como imunogênico e, na Tabela 9 aquelas que mais contribuem negativamente com esta classificação.

Tabela 8- Variáveis que mais contribuem para a classificação dos epítopos como imunogênicos

Identificador	Variáveis	Odds
D11	CHAM820101.lag2	1,190968
D12	CHOC760101.lag5	1,124578
D7	AA5.CUFQ3	1,105855
D6	AA5.MICC010101	1,099319

Tabela 9- Variáveis que mais contribuem positivamente para a classificação dos epítopos como não imunogênicos

Identificador	Variáveis	Odds
D11	BHAR880101.lag4	0,819446
	CHOC760101.lag7	0,886355
D2	TG	0,867673
D12	CIDH920105.lag5	0,92425

Verifica-se que as duas primeiras variáveis que mais contribuem com a classificação do peptídeo como imunogênico se referem à distribuição das propriedades (polarizabilidade e

área de superfície dos resíduos) ao longo da cadeia do peptídeo, sendo que a área de superfície se refere ao *lag* 5, ou seja, à posição central. O mesmo acontece para a variável CUFQ3, que se trata da classificação da densidade eletrônica da molécula e da MICC010101, que se refere à energia de ligação entre pares de aminoácidos. Ambas também são dadas para o aminoácido na posição 5, que seria aquela de contato entre o peptídeo e o receptor da célula TCD8+. Este aspecto é muito interessante, já que estudos anteriores indicaram que apenas a ligação entre o peptídeo e a molécula de MHC I não garante a sua imunogenicidade, sendo necessária também a interação com os receptores dos linfócitos, que foi aqui observado.

5.2 Modelos de Aprendizado de Máquinas para Predição da Imunogenicidade dos Epítopos

Com o objetivo de gerar um modelo preditor de imunogenicidade capaz de diferenciar os epítopos imunogênicos e não imunogênicos a partir características físicas e estruturais definidas pelas sequências primárias de aminoácidos, diferentes modelos de aprendizado de máquina foram construídos empregando métodos *ensemble* (Florestas Aleatórias e XGBoost) e o método de Máquinas de Vetores de Suporte. Estes modelos foram ajustados de modo isolado, a partir do banco de dados com todas as variáveis inicialmente propostas como potencialmente preditoras da imunogenicidade dos epítopos.

A Figura 3 apresenta os *boxplots* referentes aos valores de AUC e sensibilidade para os modelos construídos com os diferentes métodos de Aprendizado de Máquina, considerando-se as 1 000 simulações *boostrap*, nas condições de conjunto de dados completo e apenas com as variáveis selecionadas com o método *Group*-Lasso. É possível observar, para todos os métodos, em geral, uma distribuição homogênea dos valores de AUC e sensibilidade ao redor da mediana e presença de um pequeno número de *outliers*. O número de *outliers* foi maior quando avaliada a sensibilidade.

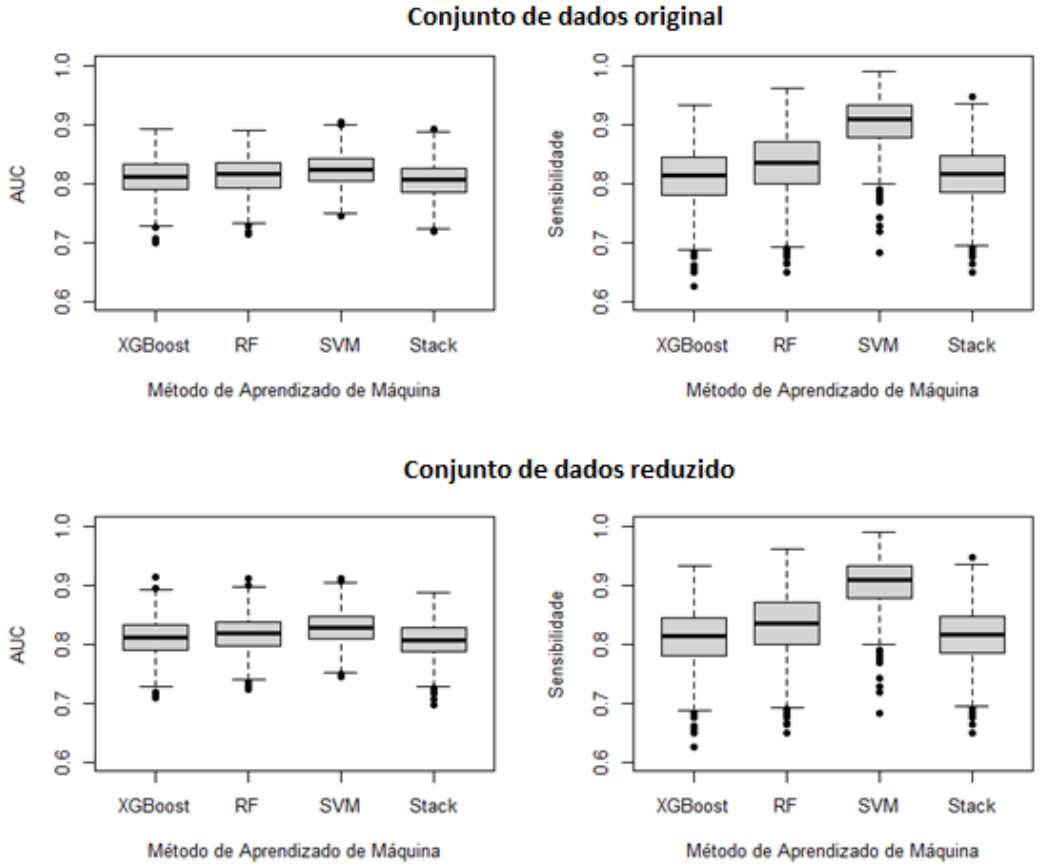

Figura 3- *Boxplots* referentes aos valores de AUC e sensibilidade para os modelos construídos com os diferentes métodos de Aprendizado de Máquina[11], considerando-se as 1.000 simulações *boostrap,*nas condições de conjunto de dados original e conjunto de dados reduzido, apenas com as variáveis selecionadas com o método *Group*-Lasso

A Tabela 10 apresenta os valores da mediana dos valores obtidos para AUC, sensibilidade e especificidade para os modelos para predição de imunogenicidade construídos.

Tabela 10-Sumário das principais métricas de desempenho do grupo de modelos gerados com as diferentes combinações de variáveis e os métodos de Aprendizado de Máquinas

Modelos	AUC	Especificidade	Sensibilidade
XGBoost	0,8108	0,8066	0,8119
***Group* Lasso + XGBoost**	0,8119	0,8137	0,8137
***Group* Lasso + SVM**	0,8230	0,7885	0,8600
***Group* Lasso + SVM**	0,8286	0,7553	0,9078
RF	0,8146	0,8040	0,8265
Group Lasso + RF	0,8176	0,8052	0,8351
Stack Model	0,8067	0,8000	0,8146
Group* Lasso + *Stack Model	0,8072	0,8000	0,8165

[11] XGBoost, RF (Florestas Aleatórias), SVM (Máquina de Vetores de Suporte), Stack (Stack model).

A análise da Figura 3 e da Tabela 10 permite verificar que, para o conjunto de dados analisado, tanto para a condição de ajuste do modelo com o *conjunto de dados* completo quanto para a condição reduzido, todos os métodos de Aprendizado Estatístico empregados (*XGBoost*, RF, SVM e Stack) conduziram a valores de medianas de AUC e sensibilidade acima de 80%, indicando uma boa capacidade para a predição da imunogenicidade de epítopos tumorais.

Quando comparados os desempenhos dos modelos construídos individualmente com os diferentes métodos (*XGBoost*, RF e SVM), se verifica que os modelos gerados com o SVM foram aqueles que apresentaram o melhor desempenho. Observa-se que, para o *conjunto de dados* original, os valores das medianas, para AUC e sensibilidade foram, respectivamente, 0,8230 e 0,8600. Para o *conjunto de dados* reduzido, os estes valores foram 0,8286 e 0,9078. Os resultados alcançados com os preditores aqui propostos são animadores e superiores aos apresentados, por exemplo, por (DIAO et al., 2022). Os autores propuseram um modelo para predição da imunogenicidade de neoantígenos utilizando redes neurais convolucionais e variáveis relacionadas à afinidade de ligação peptídeo-MCH e eficiência do transporte do peptídeo dentro da célula. No trabalho, foi observada uma acurácia de 0,801 e sensibilidade de 0,783. Os atores também verificaram o desempenho de modelos construídos com as mesmas variáveis e o método SVM. Neste caso, o resultado obtido foi de 0,778 para AUC e 0,687 para sensibilidade.

Os resultados do presente trabalho também se mostram superiores àqueles obtidos por Tung e Ho (2007) que propuseram um dos primeiros preditores de imunogenicidade de epítopos, construído com SVM e a partir de propensões físico-químicas dos aminoácidos como variáveis preditoras (TUNG; HO, 2007); e Zhang et al. (2015), que construíram um preditor de imunogenicidade de epítopos a partir de 18 grandes conjuntos de características e utilizaram algoritmo genético para seleção do melhor subconjunto de variáveis e florestas aleatórias como classificador. No estudo de Zhang, o treinamento e validação do modelo não foi realizado com epítopos tumorais e o melhor valor de AUC alcançado foi 0,846 e sensibilidade 0,715.

Os modelos construídos com o XGBoost, por outro lado, foram aqueles que apresentaram os menores valores de AUC e sensibilidade. Os valores, para estas métricas foram, respectivamente 0,8119 e 0,8119 para a base de dados original e 0,8119 e 0,8137 para os dados referentes às variáveis selecionadas na primeira etapa deste trabalho. Embora o modelo construído com o XGBoost tenha apresentado o resultado mais modesto, verifica-se que o valor de AUC alcançado é superior a trabalhos anteriores, como por exemplo, o de Zhou

et al. (2019), que construíram um modelo preditor de imunogenicidade de peptídeos utilizando os escores de hidrofobicidade como variáveis preditoras e o XGBoost. O valor de AUC alcançado pelos autores foi de 0,64, inferior àquele observado no presente trabalho que incluiu, além das propensões físico-químicas dos aminoácidos, também variáveis que tem potencial de predição da conformação estrutural dos peptídeos e, portanto, de seu reconhecimento e ligação com os receptores dos linfócitos TCD8+.

É interessante observar que a seleção de variáveis com o método de regressão logística e penalização pelo *Group*-Lasso, com redução do número total de variáveis de 5.913 para 490, conduziu a pequenos incrementos nos valores das métricas de desempenho avaliadas. Este resultado indica, portanto, que a seleção de variáveis é uma estratégia interessante para a construção de modelos preditivos menos complexos e, ao mesmo tempo, com boa capacidade de predição. Além de permitir uma maior generalização dos modelos, a redução de variáveis também está associada a um melhor potencial de interpretação e custos operacionais para caracterização de novos epítopos.

De acordo com Zhou (2012), o método ensamble do tipo *stack* permite que seja construído um meta-preditor a partir diferentes métodos que são integrados em etapas, como *first* e *second learners* (ZHOU, 2012). Neste estudo, quando considerado os modelos construídos com o método *Stack,* em comparação com os métodos de aprendizado individuais, a combinação de modelos conduziu a menores valores de mediana para AUC e sensibilidade com 1 000 corridas de *boostrap*. Os valores, para estas métricas foram, respectivamente 0,8067 e 0,8146 para a base de dados original e 0,8072 e 0,8165 para os dados referentes às variáveis selecionadas na primeira etapa deste trabalho. Ao contrário do esperado, a evidência aqui gerada mostra que o método com ajustes individuais dos modelos XGBoost e SVM como *first learners* e utilização das predições como entrada para o *second learner* Florestas Aleatória não foi efetivo para geração de um meta-preditor com maior capacidade preditiva. Este resultado pode estar associado a alguns fatores. Um deles é o fato de que, neste trabalho, embora tenha sido implementado o processo de reamostragem por *boostrap*, o conjunto de dados utilizado para as predições com os *first learners* foi o mesmo daquele utilizado para o *second learner*. Isso Pode conduzir a um *overfitting* do modelo (ZHOU, 2012). Outro fator pode ter sido a escolha da Floresta Aleatória como *second learner*. Considerando-se os resultados alcançados com os preditores *XGBoost*, RF e SVM, para os quais se verificou que o SVM foi aquele com melhor desempenho, talvez o seu posicionamento como *second learner* pudesse ter conduzido a uma melhor capacidade preditiva. Um ponto importante, entretanto, é que, embora o meta-preditor aqui proposto tenha apresentado um desempenho diferente daquele esperado, verifica-se que os valores alcançados foram superiores àqueles

apresentados, por exemplo, por Khanna e Rana (2020), que, a partir de 29 propriedades físico-químicas de aminoácidos, construíram um modelo preditivo de antigenicidade de epítopos com a combinação de Florestas Aleatórias, SVM e Florestas Aleatórias Randomizadas como *first learners*, *Blackboost* and *avNNet* como modelos intermediários e, novamente *Blackboost* and *avNNet* como *second learners*.. Com este modelo, os autores alcançaram uma AUC de 76% (KHANNA; RANA, 2020).

6 CONCLUSÕES

A predição da imunogenicidade de neoepítopos é um processo fundamental para o desenvolvimento de novos imunoterápicos para o tratamento do câncer, uma vez que seus resultados podem direcionar mais assertivamente os experimentos em bancada que, em geral, envolvem elevados tempo e custos. Neste trabalho, foi proposta a construção de diferentes modelos preditivos de imunogenicidade de epítopos, utilizando abordagens de aprendizado de máquina e 16 grandes conjuntos de variáveis potencialmente preditoras da indução da resposta imune.

A utilização do método de regressão logística com as penalizações LASSO e *Group*-LASSO conduziram à modelos com um menor número de variáveis preditivas (59 e 490, respectivamente), sendo que o modelo que considerou o agrupamento de variáveis foi aquele com melhor desempenho (AUC=0,6958). As variáveis obtidas com o método *Group*-LASSO estavam relacionadas tanto às propensões físico-químicas, características evolucionárias e estruturais dos peptídeos, sendo que aquelas que mais contribuíram positivamente com a classificação imunogênica foram aquela relacionadas aos aminoácidos do centro da cadeia.

Considerando-se os modelos construídos com as diferentes abordagens de Aprendizado de Máquinas (*XGBoost*, RF, SVM e Stack), verificou-se que a seleção prévia de variáveis resultou em alguma melhoria das medidas de desempenho dos preditores, como AUC e sensibilidade.

O modelo construído com as variáveis selecionadas com o método Group-Lasso e a abordagem SVM foi aquele com melhor desempenho. Para o conjunto de teste, foi alcançado um valor de AUC igual a 0,8286 e sensibilidade 0,9078. Este resultado se mostra superior aos demais modelos do estado da arte, indicando o grande potencial do preditor aqui desenvolvido para auxiliar o processo de desenvolvimento de imunoterápicos contra o câncer. Uma vez que os resultados foram animadores, novos estudos deverão ser conduzidos ampliando o número de peptídeos no conjunto de treinamento e estabelecendo um conjunto teste além do conjunto de validação aqui utilizado. Além disso, a construção do modelo *Stack* deverá ser revista, de modo que se alcance um meta-preditor mais robusto.

REFERÊNCIAS BIBLIOGRÁFICAS

AAindex: Amino acid index database. Disponível em: <https://www.genome.jp/aaindex/>. Acesso em: 16 abr. 2023.

ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. **Cellular and molecular immunology**. Eighth edition ed. Philadelphia, PA: Elsevier Saunders, 2015.

ADAMCZAK, R.; POROLLO, A.; MELLER, J. Combining prediction of secondary structure and solvent accessibility in proteins. **Proteins: Structure, Function, and Bioinformatics**, v. 59, n. 3, p. 467–475, 2005.

AL-BARAKATI, H. J. et al. RF-GlutarySite: Random Forest based predictor for Glutarylation sites. [s.d.].

ALBERTS et al. **Biologia Molecular da Célula**. 6. ed. Porto Alegre: ARTMED EDITORA LTDA, 2017.

BORRA, S.; DI CIACCIO, A. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. **Computational Statistics & Data Analysis**, v. 54, n. 12, p. 2976–2989, 1 dez. 2010.

BUNZ, F. **Principles of cancer genetics**. Third edition ed. Cham: Springer, 2022.

CAI, C. Z. et al. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. **Nucleic Acids Research**, v. 31, n. 13, p. 3692–3697, 1 jul. 2003.

CAI, Y. et al. Artificial intelligence applied in neoantigen identification facilitates personalized cancer immunotherapy. **Frontiers in Oncology**, v. 12, p. 1054231, 2022.

CASTRO, A. et al. Subcellular location of source proteins improves prediction of neoantigens for immunotherapy. **The EMBO journal**, v. 41, n. 24, p. e111071, 15 dez. 2022.

CHEN, C. et al. Using pseudo-amino acid composition and support vector machine to predict protein structural class. **Journal of Theoretical Biology**, v. 243, n. 3, p. 444–448, 7 dez. 2006.

CHEN, C. et al. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. **Computers in Biology and Medicine**, v. 123, p. 103899, 1 ago. 2020.

CHEN, J. et al. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. **Amino Acids**, v. 33, n. 3, p. 423–428, 1 set. 2007.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**... Em: KDD '16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. San Francisco California USA: ACM, 13 ago. 2016. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 22 abr. 2023

CHOU, K. C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. **Biochemical and Biophysical Research Communications**, v. 278, n. 2, p. 477–483, 19 nov. 2000.

CHOU, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. **Proteins: Structure, Function, and Bioinformatics**, v. 43, n. 3, p. 246–255, 2001.

CHOU, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. **Bioinformatics**, v. 21, n. 1, p. 10–19, 1 jan. 2005.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1 set. 1995.

CUI, J. et al. Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. **Molecular Immunology**, v. 44, n. 5, p. 866–877, 1 fev. 2007.

DIAO, K. et al. Seq2Neo: A Comprehensive Pipeline for Cancer Neoantigen Immunogenicity Prediction. **International Journal of Molecular Sciences**, v. 23, n. 19, p. 11624, jan. 2022.

DUBCHAK, I. et al. Prediction of protein folding class using global description of amino acid sequence. **Proceedings of the National Academy of Sciences**, v. 92, n. 19, p. 8700–8704, 12 set. 1995.

Estimativa 2023: incidência de câncer no Brasil | INCA - Instituto Nacional de Câncer. Disponível em: <https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-no-brasil>. Acesso em: 27 abr. 2023.

FOTAKIS, G.; TRAJANOSKI, Z.; RIEDER, D. Computational cancer neoantigen prediction: current status and recent advances. **Immuno-Oncology Technology**, v. 12, p. 100052, dez. 2021.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, out. 2001.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. **Journal of statistical software**, v. 33, n. 1, p. 1–22, 2010.

GARCIA-GARIJO, A.; FAJARDO, C. A.; GROS, A. Determinants for Neoantigen Identification. **Frontiers in Immunology**, v. 10, p. 1392, 24 jun. 2019.

GOVINDAN, G.; NAIR, A. S. **Composition, Transition and Distribution (CTD) — A dynamic feature for predictions based on hierarchical structure of cellular sorting**. 2011 Annual IEEE India Conference. **Anais**... Em: 2011 ANNUAL IEEE INDIA CONFERENCE. dez. 2011.

HANAHAN, D. Hallmarks of Cancer: New Dimensions. **Cancer Discovery**, v. 12, n. 1, p. 31–46, 1 jan. 2022.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. 2nd ed ed. New York, NY: Springer, 2009a.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. 2nd ed ed. New York, NY: Springer, 2009b.

HEMMATEENEJAD, B.; YOUSEFINEJAD, S.; MEHDIPOUR, A. R. Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides. **Amino Acids**, v. 40, n. 4, p. 1169–1183, 1 abr. 2011.

HENIKOFF, S.; HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 22, p. 10915–10919, 15 nov. 1992.

HU, Z.; OTT, P. A.; WU, C. J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. **Nature Reviews Immunology**, v. 18, n. 3, p. 168–182, mar. 2018.

JAMES, G. et al. (EDS.). **An introduction to statistical learning: with applications in R**. New York: Springer, 2013.

JIANG, T. et al. Tumor neoantigens: from basic research to clinical applications. **Journal of Hematology & Oncology**, v. 12, n. 1, p. 93, 6 set. 2019.

KHANNA, D.; RANA, P. S. Improvement in prediction of antigenic epitopes using stacked generalisation: an ensemble approach. **IET Systems Biology**, v. 14, n. 1, p. 1–7, 1 fev. 2020.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. **Anais**...: IJCAI'95.San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 20 ago. 1995. . Acesso em: 14 abr. 2023

KOUROU, K. et al. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. **Computational and Structural Biotechnology Journal**, v. 19, p. 5546–5555, 6 out. 2021.

KRINGELUM, J. V. et al. Structural analysis of B-cell epitopes in antibody:protein complexes. **Molecular Immunology**, v. 53, n. 1, p. 24–34, 1 jan. 2013.

LI, Q. et al. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. **Journal of Translational Medicine**, v. 20, n. 1, p. 177, 18 abr. 2022.

LIMONGELLI, I.; MARINI, S.; BELLAZZI, R. PaPI: pseudo amino acid composition to score human protein-coding variants. **BMC Bioinformatics**, v. 16, n. 1, p. 123, 19 abr. 2015.

LIU, J. et al. Cancer vaccines as promising immuno-therapeutics: platforms and current progress. **Journal of Hematology & Oncology**, v. 15, n. 1, p. 28, 18 mar. 2022.

LIU, T.; SHI, K.; LI, W. Deep learning methods improve linear B-cell epitope prediction. **BioData Mining**, v. 13, n. 1, p. 1, 17 abr. 2020.

MEIER, L.; VAN DE GEER, S.; BÜHLMANN, P. The Group Lasso for Logistic Regression. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 70, n. 1, p. 53–71, 1 fev. 2008.

ONG, S. A. et al. Efficacy of different protein descriptors in predicting protein functional families. **BMC Bioinformatics**, v. 8, n. 1, p. 300, 17 ago. 2007.

O'SHEA, R. J. et al. Sparse Regression in Cancer Genomics: Comparing Variable Selection and Predictions in Real World Data. **Cancer Informatics**, v. 20, p. 11769351211056298, 27 nov. 2021.

PARK, K.-J.; KANEHISA, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. **Bioinformatics (Oxford, England)**, v. 19, n. 13, p. 1656–1663, 1 set. 2003.

RIBATTI, D. The concept of immune surveillance against tumors. The first theories. **Oncotarget**, v. 8, n. 4, p. 7175–7180, 24 jan. 2017.

RILEY, R. S. et al. Delivery technologies for cancer immunotherapy. **Nature reviews. Drug discovery**, v. 18, n. 3, p. 175–196, mar. 2019.

SAETHANG, T. et al. PAAQD: Predicting immunogenicity of MHC class I binding peptides using amino acid pairwise contact potentials and quantum topological molecular similarity descriptors. **Journal of Immunological Methods**, v. 387, n. 1, p. 293–302, 31 jan. 2013.

SCHUELER-FURMAN, O. et al. Structure-based prediction of binding peptides to MHC class I molecules: Application to a broad range of MHC alleles. **Protein Science**, v. 9, n. 9, p. 1838–1846, 2000.

TIAN, Y.; CHEN, L.; JIANG, Y. LASSO-based screening for potential prognostic biomarkers associated with glioblastoma. **Frontiers in Oncology**, v. 12, p. 1057383, 16 jan. 2023.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

TIEN, M. Z. et al. Maximum Allowed Solvent Accessibilites of Residues in Proteins. **PLoS ONE**, v. 8, n. 11, p. e80635, 21 nov. 2013.

TOTH, R. et al. Random forest-based modelling to detect biomarkers for prostate cancer progression. **Clinical Epigenetics**, v. 11, n. 1, p. 148, 22 out. 2019.

TSENG, Y.-J. et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. **International Journal of Medical Informatics**, v. 128, p. 79–86, 1 ago. 2019.

TUNG, C.-W. et al. POPISK: T-cell reactivity prediction using support vector machines and string kernels. **BMC Bioinformatics**, v. 12, n. 1, p. 446, 15 nov. 2011.

TUNG, C.-W.; HO, S.-Y. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. **Bioinformatics**, v. 23, n. 8, p. 942–949, 15 abr. 2007.

WALDMAN, A. D.; FRITZ, J. M.; LENARDO, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. **Nature Reviews Immunology**, v. 20, n. 11, p. 651–668, nov. 2020.

XIE, C.; SHI, Y.; ZHANG, C. **Deeplearning based MHC epitope prediction for cancer neoantigen discover**. , 2021. Disponível em: <https://doi.org/10.1101/2021.11.10.468160>. Acesso em: 27 abr. 2023

XIE, N. et al. Neoantigens: promising targets for cancer therapy. **Signal Transduction and Targeted Therapy**, v. 8, p. 9, 6 jan. 2023.

XU, Y.; GOODACRE, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. **Journal of Analysis and Testing**, v. 2, n. 3, p. 249–262, 2018.

YANG, Y.; ZOU, H. A fast unified algorithm for solving group-lasso penalize learning problems. **Statistics and Computing**, v. 25, n. 6, p. 1129–1141, nov. 2015.

YU, S.-H. et al. LASSO and Bioinformatics Analysis in the Identification of Key Genes for Prognostic Genes of Gynecologic Cancer. **Journal of Personalized Medicine**, v. 11, n. 11, p. 1177, 11 nov. 2021.

ZHANG, W. et al. Computational Prediction of Conformational B-Cell Epitopes from Antigen Primary Structures by Ensemble Learning. **PLOS ONE**, v. 7, n. 8, p. e43575, 21 ago. 2012.

ZHANG, W. et al. Accurate Prediction of Immunogenic T-Cell Epitopes from Epitope Sequences Using the Genetic Algorithm-Based Ensemble Learning. **PLOS ONE**, v. 10, n. 5, p. e0128194, 28 maio 2015.

ZHANG, Y.; ZHANG, Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. **Cellular & Molecular Immunology**, v. 17, n. 8, p. 807–821, ago. 2020.

ZHOU, C. et al. pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. **Genome Medicine**, v. 11, n. 1, p. 67, 30 out. 2019.

ZHOU, Z.-H. **Ensemble methods: foundations and algorithms**. Boca Raton, FL: Taylor & Francis, 2012.