UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE GRADUAÇÃO
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

CLEITON FELIPE VALANDRO

PERFIL DE METILAÇÃO DO GENE NT5E EM AMOSTRAS DE TUMORES SÓLIDOS BASEADO EM FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL

SÃO LEOPOLDO
2021

CLEITON FELIPE VALANDRO

PERFIL DE METILAÇÃO DO GENE NT5E EM AMOSTRAS DE TUMORES SÓLIDOS BASEADO EM FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL

Artigo apresentado como requisito parcial para obtenção do título de Bacharel em 2022, pelo Curso de Sistemas de informação da Universidade do Vale do Rio dos Sinos - UNISINOS

Orientador: Cristiano André da Costa

SÃO LEOPOLDO

2021

PERFIL DE METILAÇÃO DO GENE NT5E EM AMOSTRAS DE TUMORES SÓLIDOS BASEADO EM FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL

Cleiton Felipe Valandro
Cristiano André da Costa

O desenvolvimento deste software tem como motivação a dificuldade em caracterização do perfil de determinados genes em diferentes tipos de câncer, pois em muitas vezes é investido grandes quantidades de recursos para a realização de exames e a indefinição ainda persiste, trazendo gastos desnecessários de recursos e um tratamento tardio, reduzindo as chances de tratamento ao paciente. Será apresentado ao longo desse artigo, todo o processo, detalhamento e conclusão do desenvolvimento de um software que possui como finalidade, o auxílio a profissionais ligados à saúde, sendo eles médicos ou pesquisadores, a terem a partir deste software, mais uma ferramenta à disposição. A partir de um modelo de predição, espera-se contribuir para eventuais momentos em que os exames convencionais não sejam suficientes para a identificação de determinado tecido ser ou não ser um tecido tumoral, ou ter um pior prognóstico. O software em questão, fez uso de inteligência artificial, utilizando a técnica de Florestas Aleatórias com os dados extraídos da base de dados do The Cancer Genome Atlas (TCGA), com o objetivo de adquirir um valor de predição da expressão do gene NT5E em conjunto com os sítios de metilação. Como resultado, foram gerados métricas em forma de mapa de calor com valores de Acurácia, Precisão, Recall e F1-Score, utilizando modelo baseado em aprendizado de máquina em um ambiente colaborativo.

Palavras-chave: TCGA, *NT5E*/CD73, Aprendizado de máquina e inteligência artificial.

1 INTRODUÇÃO

Através de informações obtidas em base de dados públicas e abertas para pesquisa como o The Cancer Genome Atlas (TCGA), é possível encontrar uma grande quantidade de informações genômicas e epigenômicas (Cooper, Demicco et al. 2018), mas que são de difícil manuseio pois exigem do usuário conhecimentos relacionados a tecnologias para a extração dos dados. Nesse contexto, nota-se a necessidade de desenvolver um software com esta finalidade, pois através de interfaces intuitivas a análise de dados destas plataformas se torna mais simples e acessível para a obtenção de métodos preditivos.

Em razão da dificuldade em criar um modelo preditivo para caracterizar um perfil de determinados genes em diferentes tipos de câncer e de grandes investimentos que podem gerar algum resultado inconclusivo, o objetivo deste artigo é gerar um modelo que permita diferenciar amostras normais e tumorais a partir da expressão e dos níveis de metilação do gene *NT5E*. Neste sentido, foi desenvolvido um software com a finalidade de classificar as amostras de pacientes utilizando aprendizagem de máquina. Em decorrência a alguns tipos tumorais possuírem baixas quantidades de amostras, notou-se que o modelo de Florestas Aleatórias (Random Forest), além de poder ser utilizado em diferentes tipos de previsões, demonstra ter bons resultados quando aplicado em baixas quantidades de amostras, sendo também simples e fácil de utilizar (Tyralis and Papacharalampous, 2017).

Os dados utilizados foram baixados no formato .csv do TCGA-PANCANCER, através da ferramenta Xena Browser. A leitura dos arquivos .csv e a estatística descritiva foram feitas através da biblioteca pandas, os gráficos foram elaborados pela biblioteca Numpy e o modelo de aprendizado de máquina foi performado pela biblioteca Scikit-Learn.

Mesmo com a existência de muitos trabalhos relacionados na área, ainda é pouco explorado o mecanismo de desenvolvimento de software próprio, ou um tipo de software que seja fácil de utilizar e expansível para outros objetivos similares. Este artigo, proporciona um software livre de forma que outras pessoas da área possam implementar novas funcionalidades com o objetivo de facilitar o acesso e manuseio de pessoas que possuem pouco ou nenhuma familiaridade com desenvolvimento de software.

Para o desenvolvimento deste artigo, teremos ao longo deste, informações que demonstram a importância da utilização de ambientes colaborativos para o desenvolvimento de software e qual a linguagem de programação está atualmente mais preparada para receber modelos de aprendizado de máquina e análise de dados. Tendo em vista estas informações, o software proporciona o acesso a análises e estatísticas e como resultado, um modelo que possa diferenciar as amostras normais de amostrar tumorais, através de métricas que se tornam fundamentais na identificação de quais tipos tumorais e análises tiveram o melhor desempenho.

2 FUNDAMENTAÇÃO TEÓRICA

O diagnóstico e avaliação da progressão tumoral atualmente é feita através de diferentes tipos de exames, como por exemplo o exames de moleculares e de imagem, onde em combinação com outros exames de rotina, podem no final, custar um valor bem elevado (Zheng and Xu, 2020).

Através de informações obtidas da base de dados do The Cancer Genome Atlas (TCGA) é possível explorar os impactos das alterações genômicas nas metilações de DNA (Spainhour, Lim et al. 2019). Notou-se que a metilação do DNA é importante na regulação do gene, o que tem interferência em diversos tipos de câncer, dentro da análise de Pân-cancer, é possível obter os dados com o objetivo de analisar e identificar os padrões entre as metilações de DNA e a expressão do gene (Spainhour, Lim et al. 2019).

A utilização de ferramentas preditivas que considerem o pareamento das amostras, ou seja, tecido tumoral e seu respectivo tecido adjacente não tumoral do mesmo paciente, pode ter grande valia na avaliação da progressão tumoral e diagnóstico uma vez que é possível detectar padrões entre as amostras normal e a tumoral e a partir deste, buscar-se então uma terapia personalizada e mais efetiva para cada paciente (Aran, Camarda et al. 2017).

Neste momento, há um grande acervo de modelos de aprendizado de máquina disponíveis para uso, onde é possível destacar a utilização do algoritmo de Florestas Aleatórias como forma de obter um valor preditivo, pois o mesmo tem obtido bons resultados quando aplicado em baixas quantidades de amostras (Tyralis and Papacharalampous, 2017) e através do Python, há bibliotecas voltadas para

este modelo de forma bem simplificadas, tornando o uso do mesmo algo muito fácil de ser manuseado e desenvolvido.

Na sequência, será demonstrado com um maior nível de detalhamento, as informações de cada etapa fundamental apresentada neste artigo, desde informações sobre o perfil de metilação do gene NT5E, uso de ambientes compartilhados, utilização da linguagem Python, análises, aplicações no modelo de Florestas Aleatórias e por fim, as métricas resultantes do modelo de aprendizado de máquina utilizado.

2.1 Perfil de metilação do gene NT5E

A importância do gene *NT5E* em diversos tipos tumorais tem sido explorada em diversos estudos nos últimos anos, conforme revisado recentemente por Stagg e colaboradores (Allard, Allard et al. 2020). Devido a sua relevância no prognóstico destes diversos tipos tumorais, pelo menos sete estudos dos clínicos atualmente buscam avaliar se a inibição deste gene pode inibir o crescimento tumoral e a capacidade de formar metástase para outros tecidos (Allard, Allard et al. 2020).

A metilação do DNA é um dos mecanismos responsáveis por modular a expressão de diversos genes em condições fisiológicas ou em condições patológicas, incluindo tumores malignos e benignos, e pode aumentar ou diminuir a expressão de um determinado gene (Casalino and Verde, 2020). Embora pobremente explorado, para a regulação do gene *NT5E*, a metilação se mostrou ter efeitos no câncer de mama (Lo Nigro, Monteverde et al. 2012), em melanomas primários e metastáticos (Wang, Lee et al. 2012, Jeong, Oh et al. 2020) e em câncer de pâncreas (Chen, Pu et al. 2020).

Ao nosso conhecimento, nenhum estudo foi publicado até momento comparando o perfil de metilação do gene *NT5E* entre os diversos tipos tumorais. Ainda, uma revisão recente demonstra que os mecanismos de regulação gênica, como a metilação por exemplo, são subestimados no ambiente tumoral e que se há necessidade de estudo para um melhor conhecimento desta regulação nos diferentes tipos tumorais (Alcedo, Bowser et al. 2021).

2.2 Ambientes colaborativos e processamento de dados em nuvem

Os ambientes colaborativos ou mais especificamente o Google Colaboratory (Colab), são de extrema importância para o meio acadêmico multidisciplinar pois além de disponibilizar um ambiente adequado para o desenvolvimento de software ele também nos fornecem ferramentas que auxiliam no desenvolvimento e execução de código fonte que originalmente são executados e interpretados dentro de um servidor, neste ambiente os mesmos podem ser executados dentro de um simples navegador web por qualquer usuário na internet.

O Google Colaboratory (Colab) que será o grande anfitrião deste trabalho, nos proporcionara o desenvolvimento deste software onde poderá ser utilizado dentro desta ferramenta com a adoção de ferramentas de apoio que visa facilitar o desenvolvimento e manuseio do código fonte para sua finalidade.

Além da facilidade de acesso, outra característica muito importante é a facilidade de compartilhamento de informações para inúmeros usuários, além deste aspecto, podemos também frisar que o mesmo utiliza recursos computacionais próprios, sem a necessidade de qualquer configuração prévia e também disponibiliza a acesso gratuito à GPUs que são responsáveis por realizar o processamento do código fonte e através destes recursos, o usuário final não precisará obter recursos próprios e avançados para a utilização do software disponibilizado.

Os ambientes em nuvem, assim como o Google Colaboratory (Colab), nos fornecem ambientes e um acervo computacional disponibilizado em formato de serviço, onde cada usuário consome recursos computacionais mediante a contratação de serviço e a maioria das vezes isso é disponibilizado mediante pagamento de algum valor que podem ser cobradas de acordo com as regras das plataformas, mas o Google Colab disponibiliza os mesmos de forma gratuita (Kanani and Padole, 2019), facilitado o acesso de inúmeros usuários.

A diferença entre os serviços em nuvem e a computação tradicional é diretamente ligada ao compartilhamento de infraestrutura e processamento, pois em uma circunstância convencional o usuário final necessitaria obter um acervo computacional próprio, necessitando de um alto valor para a aquisição de hardware e os meios em nuvem são disponibilizados aos usuários de forma compartilhada, sendo mais eficiente e diminuindo o tempo ocioso de processamento em relação ao modelo convencional.

A utilização do Google Colab para o desenvolvimento deste software se faz necessário pois é uma plataforma gratuita, possibilitando que inúmeras pessoas consigam obter acesso a recursos computacionais quando os mesmo não estiverem disponíveis(Kanani and Padole, 2019), e este meio facilita a execução do código fonte pois é necessário apenas realizar a importação das bibliotecas conforme forem sendo necessários, diferentemente do modo convencional, onde é necessário depender de instalações e configurações que limitariam a disponibilidade do software e exigem do usuário mais conhecimento técnico e recursos.

2.3 Python como Linguagem de programação

O Python é amplamente utilizado em diversas áreas e tem se notado um constante aumento em sua utilização por conta da facilidade de todo o acervo de bibliotecas que facilitam o desenvolvimento de determinados softwares, mas principalmente em softwares voltados para uso da inteligência artificial.

O Python assim como qualquer outra linguagem de programação possui suas particularidades, tendo aspectos positivos e negativos, mas neste tópico, vamos destacar os principais aspectos que levaram a utilização do Python no desenvolvimento do software que será o resultado deste artigo.

O Python propriamente dito, possui inúmeras razões que justificam a utilização em aprendizado de máquina, uma delas é a disponibilidade de utilização de bibliotecas recentes para o aprendizado de máquina e aprendizagem profunda (Raschka, Patterson et al. 2020) e a utilização do mesmo, dentro do Google Colab, o que torna fácil o compartilhamento e utilização de diversos usuários sem conhecimentos em programação. Outro motivo é a grande acervo de bibliotecas para o consumo de dados externos ou tratamento deles, o que neste aspecto facilita muito todo o processo de obtenção dos dados, análise, processamento, aprendizado de máquina e gráficos, pois conseguimos utilizar muitos meios pré-existentes, evitando um trabalho desnecessário, consequentemente, podendo ter um foco maior na análise e resultados.

2.4 Inteligência computacional para análise de dados

A tecnologia de aprendizagem de máquina é utilizada em diferentes áreas pois conseguimos através dela ensinar o sistema a diferenciar determinadas amostras e, para isso, será necessário treinar o reconhecimento de determinados padrões. Após treinamento realizado, é necessário realizar testes a fim de validar se o método de treinamento realmente funcionou (Dey, 2016). Por fim, adquirir as métricas onde ela nos informa se os dados analisados estão ou não balanceados, junto com informações como a Acurácia, Precisão, Recall e F1-Score as quais são utilizadas para verificar a eficiência em classificar os dados.

Temos a disposição centenas ou até mesmo milhares de dados para gerar algum tipo de valor a partir deles, mas em sua grande maioria esses dados se tornam e são vistos apenas como algum tipo de dado sem muito valor agregado, deixando algumas vezes de serem analisamos de uma forma mais sistêmica e útil, para este casos, podemos aplicar o uso de inteligência artifical (Dey, 2016). Entende-se que os dados isolados muitas vezes quando estão sem nenhum tratamento ou apenas em um formato inicial, tornando-se apenas dados visto de forma superficial, e para apresentar um potencial preditivo, é necessário visualizá-lo e interpretá-los considerando o conjunto completo dos dados, após aplicado algum modelo de inteligência artificial.

Os modelos supervisionados de aprendizado de máquina necessitam de um certo auxílio ou direção para conseguir atingir seu objetivo (Dey, 2016), onde consistem em obter uma quantidade de dados em pares, especificando quais amostras são tumorais e quais não são e posteriormente aplicamos o modelo de Florestas Aleatórias, sendo suficiente para que o computador consiga prever determinados padrões. Em outras palavras, quanto mais experiência ou treinamento o computador possuir maior e melhor será o modelo e o resultado obtido, como consequência, o valor preditivo fornecido será ainda mais consistente e assertivo.

2.5 Árvores aleatórias como modelo de aprendizado de máquina

O modelo de árvore aleatória é um modelo que ganhou popularidade e podem ser utilizadas para a classificação e regressão, em diferentes tipos de previsões, onde também demonstra bons resultados quando aplicado em baixas quantidades

de amostras, sendo simples e fácil utilizar o modelo (Tyralis and Papacharalampous, 2017).

Após a aplicação do modelo de aprendizado de máquina, é possível obter as métricas correspondes aos testes realizados, onde podemos citar inicialmente a Acurácia, onde ela nos informa a exatidão dos testes realizados, ou seja, ela representa um número que nos diz o percentual de acertos dos testes realizados (Sanni and Guruprasad, 2021).

A Precisão nos informa a porcentagem de verdadeiros positivos, Recall calcula o número de amostras corretas (Sanni and Guruprasad, 2021) e a pontuação F1-Score, busca em equilibro entre a Precisão e o Recall, ou seja, buscando desta forma, uma melhor avaliação do modelo (Iwendi, Bashir et al. 2020).

Na Figura 1 a seguir, é demonstrado a fórmula condizente a cada métrica gerada através do método de florestas aleatórias.

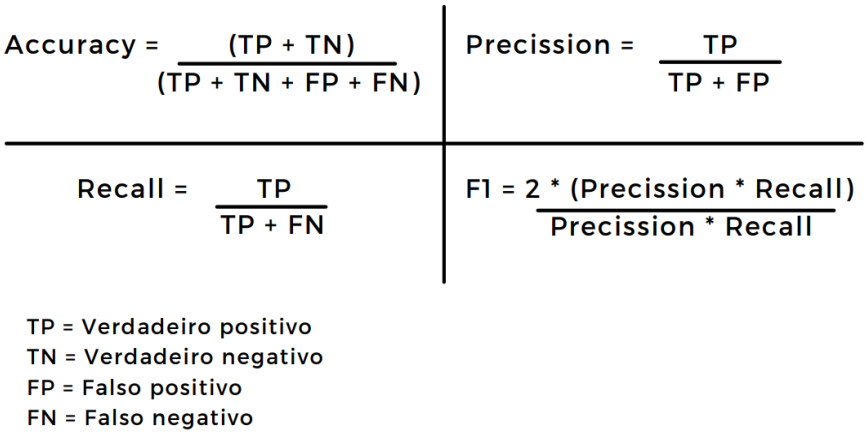

Figura 1: Resumo esquemático das métricas utilizadas no modelo. Fonte: Do autor.

3 TRABALHOS RELACIONADOS

Foram utilizados como principais referências para este artigo, cerca de três trabalhos que possuem relação ou similaridades com as questões propostas, todos os trabalhos utilizaram a plataforma do The Cancer Genome Atlas (TCGA) como base de dados e utilizando a classificação de Pan-câncer para a realização de suas análises e estabelecimento de suas conclusões.

No primeiro trabalho, foram utilizadas cerca de 9.096 amostras de 31 tipos de tumor, dados do tipo TCGA RNA-seq, sendo realizada a análise em diferentes tipos de genes, onde por fim, foi possível selecionar conjuntos de 20 genes em uma classificação correta em mais de 90% das amostras (Li, Kang et al. 2017).

No segundo trabalho, foi utilizado aprendizagem de máquina, usando o modelo de forma supervisionada, onde foi aplicado o modelo com o foco maior em câncer de mama. Os dados usados do tipo TCGA RNA-seq, contendo dados de mutações de 33 tipos diferentes de câncer para prever estados moleculares anormais em tumores (Way, Sanchez-Vega et al. 2018).

No terceiro e último trabalho, foi utilizado um modelo de aprendizado de máquina, mas com a técnica de aprendizado não supervisionado, onde foram testados a expressão de 1027 genes e como resultado foram encontrados cerca de 210 marcadores, para 74 subtipos de tumor, desta forma, abrindo caminho para novas abordagens terapêuticas (Kääriäinen, Pesola et al. 2020).

Tabela 1: Na tabela a seguir, é demonstrada uma comparação entre as diferenças e similaridades dos estudos relacionados.

	Li et al, (2017)	**Way et al, (2018)**	**Kääriäinen et al, (2020)**
Objetivo	Classificação de amostras	Classificação de amostras	Marcadores tumorais
Algoritmo	GA / KNN - Supervisionado	GA / KNN - Supervisionado	Não supervisionado
Coleta de dados	TCGA - Pan-câncer	TCGA - Pan-câncer	TCGA - Pan-câncer
Número de tipos tumorais	31	33	24
Número de amostras	9.096	9.075	6.700
Métricas	Acurácia	Precisão e Recall	Especificidade e sensibilidade

Analisando as informações dos três trabalhos relacionados, é possível notar que existem muitas similaridades e em certos pontos é apresentados um objetivo muito similar, mas com suas peculiaridades, ou seja, os três trabalhos possuem finalidades um pouco diferentes, mas os três se mostram no mesmo caminho para melhorar a identificação de tipos de tumores por meio de aprendizado de máquina e a utilização de bases de dados públicas.

É possível concluir que por mais que existam muitos trabalhos relacionados na área, ainda é pouco explorado o mecanismo de desenvolvimento de software próprio, ou seja, ou um tipo de software fácil de utilizar e expansível para outros

meios ou objetivos. Dois dos três trabalhos utilizaram mecanismos externos para a análise dos dados coletados, isso por muitas vezes, podem dificultar a utilização por outras pessoas, deste modo, o artigo em questão, busca não apenas identificar amostras tumorais, mas também proporcionar um software livre de forma que outras pessoas da área possam implementar novas funcionalidades com o objetivo de facilitar o acesso e manuseio de pessoas que possuem pouco ou nenhuma familiaridade com desenvolvimento de software.

Esse software também é necessário pois tem como foco, análises dos dados utilizando a expressão de gene *NT5E* para o câncer de tireoide mesmo tendo como resultado outros tipos tumorais além deste. Através da expressão do gene *NT5E* foi possível obter métricas e resultados na tentativa de disponibilidade um valor preditivo tanto para o valor do gene quanto para as combinações com os sítios de metilação.

4 MATERIAIS E MÉTODOS

Através do software mencionado neste artigo, se tornou possível a realização da coleta dos dados da base do The Cancer Genome Atlas (TCGA), onde foram obtidos os dados para análise de Pan-câncer, dados como expressão do gene, fenótipos e sítios de metilação, com intermédio da biblioteca XenaPython. Dentro da biblioteca XenaPython, é possível estabelecer um filtro prévio, onde selecionamos qual o gene seria utilizado, deste modo, foram importados todos os dados respectivos do gene NT5E.

Após armazenar e estruturar os dados, o próximo passo foi a seleção dos dados, onde separamos eles por pares, ou seja, foram separadas as amostras em tumoral e normal, deste modo, os dados que não possuíam pares foram alocados para outra planilha, ficando apenas na planilha original, os dados pareados. Já com os dados pareados, foi possível realizar a análise e estatística, onde posteriormente foi necessário separar e rotular os dados classificados como amostra normal e amostra tumoral para serem utilizados como treinamento e teste no modelo de aprendizado de máquina.

Na Figura 2 é possível identificar o fluxo de forma simples e resumida, onde iniciamos com a importação dos dados, pareamento, análise, estatística e aplicação no modelo de aprendizado de máquina.

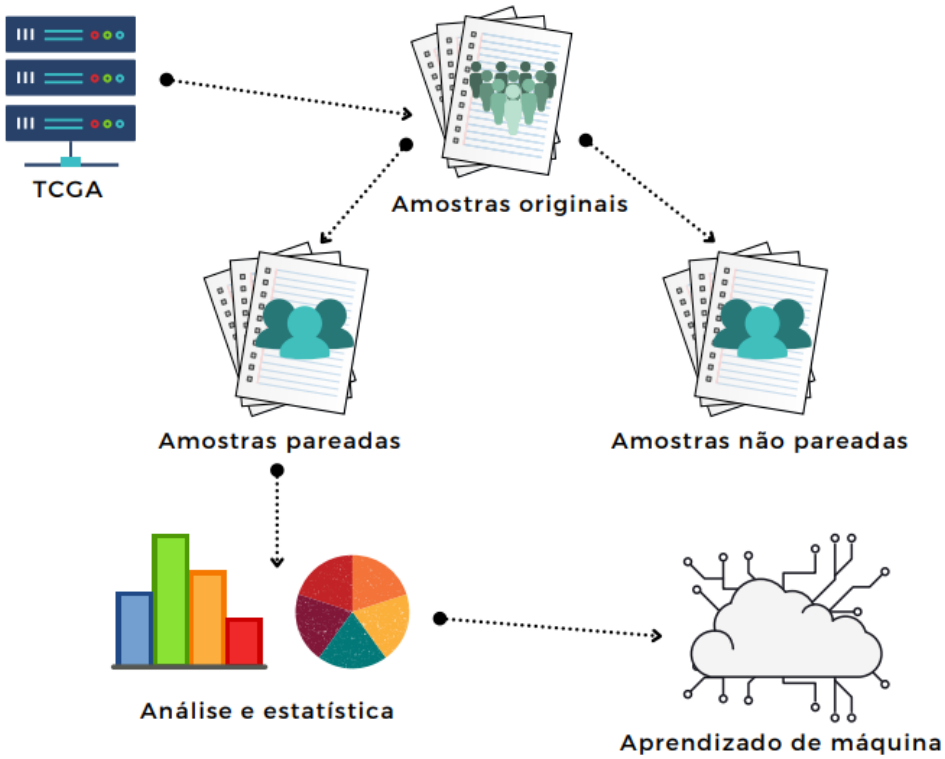

Figura 2: Fluxo esquemático utilizado para a realização do presente trabalho. Fonte: Do autor.

4.1 Base de dados

Os dados apresentados neste trabalho foram extraídos do portal *The Cancer Genome Atlas* (https://portal.gdc.cancer.gov/). Para a utilização na validação do software, foi realizado o download dos dados em agosto de 2021 em formato .csv.

4.2 Aspectos éticos

A base de dados TCGA que foram utilizadas para este estudo possuem livre acesso e o as identificações dos pacientes são codificadas. Assim, não há

necessidade de submissão a Comitê de Ética em Pesquisa com Seres Humanos desta instituição.

4.3 Ferramentas computacionais

Para este artigo, foi utilizado o ambiente colaborativo do Google Colaboratory (Colab), este ambiente é disponibilizado gratuitamente e de fácil compartilhamento, dentro deste ambiente o usuário pode contar com a utilização da linguagem de programação Python que originalmente é voltada para servidor, mas nesta ferramenta é disponibilizada em um formato mais intuitivo, podendo ser utilizada através da ferramenta Jupyter Notebook, que torna a utilização do Python diretamente no front-end do navegador.

No aprendizado de máquina foi utilizado o modelo de aprendizagem supervisionada, mais precisamente o modelo de Florestas Aleatórias a partir da biblioteca scikit-learn, antecedendo de outras bibliotecas fundamentais como Pandas que nos auxiliam no manuseio dos dados, XenaPython realizando a intermediação entre a base de dados e o Google Colab, Numpy para operações matemáticas, Plotly fez a geração de boxplots, Matplotlib disponibilizou os demais gráficos e por fim, o Seaborn gerou os mapas de calor para os valores resultantes do modelo de aprendizado de máquina.

4.3 Procedimentos e seleção dos dados

Conforme havíamos descrito nos itens anteriores, foi necessário importar da base de dados do The Cancer Genome Atlas (https://portal.gdc.cancer.gov/) cerca de 10535 amostras, destas, foram selecionados 1504 dados que existiam algum tipo de par em diferentes tipos tumorais, separando assim as amostras pareadas das amostras que não existiam pares correspondentes. Estes pares em questão, foram novamente selecionados, ou seja, foi aplicado um novo filtro, separando pares em identificadores um pouco mais específicos, sendo considerado amostras pareadas aquelas que tinham a amostra normal e a amostra tumoral correspondente, podendo ser identificada como tumoral as amostras que possuíam em seu final, o identificador (01) e o identificador (05), restando as amostras com identificador (11) para as amostras que forem consideradas como normais.

A Tabela 2 demostra quais são os tipos passíveis a se tornarem pares dentro da seleção das amostras coletadas da base de dados.

Identificador	Tipo de amostra
01	**Tumoral**
11	**Normal**
05	**Tumoral**
11	**Normal**

Tabela 2: Classificação das amostras do banco TCGA. Fonte: Do autor.

4.4 Estatística descritiva e análise estatística.

A partir das amostras pareadas, foi possível criar diversos relatórios e análises que contribuíram para a melhor identificação dos dados que foram aplicados no modelo de aprendizado de máquina para obter um melhor resultado e um valor preditivo mais preciso, contribuindo também para a visualização do usuário em uma forma mais sistêmica de todo o processo realizado e de uma melhor visualização e análise das amostras selecionadas e pareadas.

Os tipos de análises anteriormente descritas, foram disponibilizadas em forma de tabelas, gráficos, boxplot e mapa de calor, onde foi possível identificar melhor as informações das amostras como quantidade e percentual correspondentes aos grupos de tipos tumorais e de seus sítios de metilação, sendo eles os dados originais, pareados e não pareados.

Foi possível criar relatórios contendo alguns dados matemáticos que se mostram fundamentais para a definição de quais tipos tumorais possuem dados suficientes e quais tipos tumorais tiveram a probabilidade maior de obter sucesso na análise e aplicação no modelo de aprendizado de máquina, podendo citar como exemplo o valor de p, número de amostras, desvio padrão, média, mediana, valores máximos e valores mínimos.

A análise estatística dos dados de expressão do gene *NT5E* foi realizada através do Teste t de Student para dados pareados, onde foi considerado o valor menor e igual a 0.01.

4.5 Aplicação do aprendizado de máquina utilizando o modelo de Florestas aleatórias

Para a aplicação dos dados em um modelo de aprendizado de máquina, foi necessário realizar diversas análises prévias para definir quais os tipos tumorais se mostraram com a maior probabilidade de terem bons resultados, antecedendo a esta definição, foi necessário também, organizar os dados separando eles por pares, onde foi definido quais os tipos de amostras através dos identificadores, seriam utilizadas como tumoral e não tumoral, onde posteriormente, foram rotuladas cada amostra com seu respectivo diagnóstico, pois o modelo de Florestas Aleatórias utiliza aprendizado supervisionado, então, neste cenário, é necessário identificar cada amostra para que ele aprenda e consiga de fato testar e obter suas métricas resultantes.

Já com os dados devidamente separados por pares e rotulados com base no identificador de cada amostra, foi necessário definir dentro do algoritmo, quais colunas ou campos foram utilizadas dentro do modelo, pois o algoritmo precisará saber quais campos são os dados gerais que compõem o aprendizado e qual é o campo do resultado, deste modo, ele utilizou esta combinação para criar um algoritmo capaz de separar determinadas amostras, usando como teste uma porcentagem definida previamente à execução do modelo.

Os dados ou amostras pareadas foram separados aleatoriamente entre a base de treinamento e a base de testes, em um percentual predefinido pelo usuário, mas neste artigo foi demonstrado métricas referente a um percentual de 50% para treinamento e outros 50% para os testes, onde foi aplicado aprendizado de máquina no modelo de Florestas Aleatórias para os dados do gene *NT5E* separadamente e posteriormente foi aplicado o mesmo modelo para a combinação entre os valores do gene *NT5E* com cada sítio de metilação, onde foi realizado o mesmo procedimento para todos os tipos tumorais que tiverem um valor de $p \leq 0.01$.

4.6 Código fonte

A seguir, será compartilhado o código fonte utilizado ao decorrer do artigo, onde será disponibilizado tanto para acesso no Github quanto para acesso no Google Colab.

Os dois repositórios são de fácil acesso e gratuitos, onde através do Github, é possível armazenas todo o código e visioná-lo, proporcionando também que outras pessoas realizem contribuições para melhorar ou desenvolver novas funcionalidades.

Github:
https://github.com/CleitonValandro/predictive-model-in-tumor-samples-using-artificial-intelligence.git

Google Colab:
https://colab.research.google.com/drive/1T4IxBxXnk37_W2UkKG8dcHDABQ7UY-Qz?usp=sharing

5 RESULTADOS

Foram coletados 10535 amostras que continham o valor de expressão do gene *NT5E* e informações de fenótipos e sítios de metilação. As 10535 amostras que foram utilizadas como base inicial, possuíam 8 tipos de identificadores de amostras (01, 02, 03, 05, 06, 07, 11 e 20) e 33 tipos tumorais. Destas amostras coletadas, foram selecionadas inicialmente, as amostras que possuíam algum tipo de pareamento, resultando em 1504 amostras pareadas

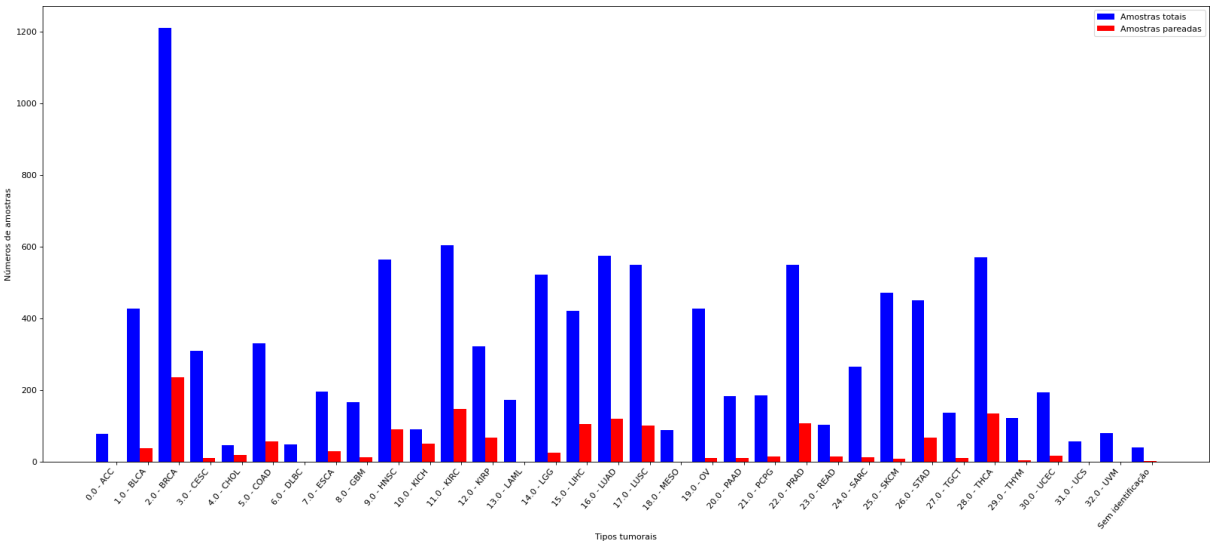

Figura 3: Gráfico comparativo em diferentes tipos tumorais demonstrando a quantidade de amostras pareadas (Vermelho) em relação à quantidade de amostras totais (Azul). Fonte: Do autor.

Tabela 3: Abreviatura e nome completo de cada tipo tumoral utilizadas neste estudo. Fonte: Do autor.

Abreviação	**Tipo Tumoral completo**
BLCA	Carcinoma urotelial da bexiga
BRCA	Carcinoma invasivo da mama
CESC	Carcinoma de células escamosas cervicais e adenocarcinoma endocervical
CHOL	Colangiocarcinoma
COAD	Adenocarcinoma do cólon
DLBC	Linfoma Linfóide Neoplasma Difuso de Grandes Células B
ESCA	Carcinoma esofágico
HNSC	Carcinoma espinocelular de cabeça e pescoço
KICH	Carcinoma cromófobo de células renais
KIRC	Carcinoma renal de células claras
KIRP	Carcinoma de células papilares renais
LGG	Glioma de grau inferior do cérebro
LIHC	Carcinoma hepatocelular de fígado
LUAD	Adenocarcinoma de pulmão
LUSC	Carcinoma de células escamosas de pulmão
MESO	Mesotelioma
OV	Cistadenocarcinoma seroso de ovário
PAAD	Adenocarcinoma pancreático
PRAD	Adenocarcinoma de próstata
READ	Adenocarcinoma de reto
SARC	Sarcoma
SKCM	Melanoma cutâneo de pele
STAD	Adenocarcinoma de estômago
TGCT	Tumores de células germinativas testiculares
THCA	Carcinoma de tireoide
THYM	Timoma
UCEC	Carcinoma Endometrial do Corpo Uterino

Após, foram aplicados novos filtros, onde foram selecionadas apenas as amostras que existiam pares do tumor (Primário e Novo primário) e seu respectivo tecido normal (Tecido Sólido Normal), resultando em um novo número de 1440 amostras.

Tabela 4: Quantidade e percentual de cada identificador em relação as amostras originais e amostras pareadas. Fonte: Do autor.

		Amostras originais		**Amostras pareadas**	
Descrição	Identificador	Quantidade	Percentual	Quantidade	Percentual
Tumoral Primário	01	9186	87.19	**748**	49.73
Tumor Recorrente	02	44	0.41	33	2.19
Sangue Periférico	03	173	1.64		
Novo Primário	05	11	0.10	**11**	0.73
Metastático	06	392	3.72	30	1.99
Metastático Adicional	07	1	0.0094	1	0.06
Tecido Sólido Normal	11	727	6.90	**681**	45.27
Analito de Controle	20	1	0.0094		

Nas 1440 amostras resultantes da nova seleção, aplicou-se o teste T de Student para avaliar se havia diferença estatística na expressão do gene NT5E entre as amostras normais e tumorais. Cada subtipo tumoral foi avaliado isoladamente. Os resultados da análise estão apresentados na Tabela 5, na qual são apresentados os seguintes dados: valor de P (P VALUE), números de amostras pareadas (N), média (mean), mediana (median), desvio padrão (sd), mínimos (min) e máximos (max), sendo antecedidos por (T) de tumoral e (NT) de não tumoral. Sete tipos tumorais não apresentaram pelo menos 3 amostras pareadas para serem consideradas no teste estatístico, sendo eles SARC, THYM, GBM, LGG, OV, SKCM e TGCT. Os tipos tumorais CESC, PAAD, COAD, ESCA, KIRP, READ e SARC não apresentaram diferença estatística para expressão do gene *NT5E*. Por fim, nove tipos tumorais, destacados em negrito na Tabela 5, apresentaram uma diferença estatística um valor $p \leq 0.01$ e estes então foram selecionados previamente para o modelo.

Tabela 5: Análise e estatística das informações geradas a partir das amostras pareadas. Fonte: Do autor.

Amostra	N	Amostras Normais					Amostras Tumorais					P VALUE
		NT_mean	NT_median	NT_sd	NT_min	NT_max	T_mean	T_median	T_sd	T_min	T_max	
2.0 - BRCA	**112**	**3,43**	**3,48**	**0,99**	**-0,01**	**5,29**	**2,45**	**2,47**	**1,05**	**-1,36**	**5,21**	**< 0,000**
9.0 - HNSC	**43**	**1,63**	**1,40**	**1,10**	**-0,36**	**4,31**	**3,27**	**3,06**	**1,71**	**-1,28**	**6,97**	**< 0,000**
16.0 - LUAD	**58**	**2,47**	**2,60**	**0,59**	**0,59**	**3,42**	**4,06**	**4,27**	**1,49**	**0,91**	**7,66**	**< 0,000**
22.0 - PRAD	**52**	**2,77**	**2,82**	**0,79**	**0,53**	**4,54**	**1,58**	**1,81**	**0,83**	**-0,62**	**3,13**	**< 0,000**
28.0 - THCA	**59**	**3,53**	**3,49**	**0,65**	**2,10**	**4,94**	**4,59**	**5,06**	**1,37**	**0,40**	**6,87**	**< 0,000**
11.0 - KIRC	**72**	**2,57**	**2,92**	**1,37**	**-1,00**	**6,04**	**3,35**	**3,47**	**0,90**	**0,77**	**5,51**	**< 0,000**
10.0 - KICH	**25**	**2,22**	**1,91**	**1,35**	**-0,32**	**4,29**	**0,60**	**0,20**	**1,80**	**-2,73**	**6,67**	**0,001**
1.0 - BLCA	**19**	**3,15**	**3,61**	**0,97**	**0,89**	**4,28**	**1,58**	**1,32**	**1,63**	**-1,94**	**5,41**	**0,001**
15.0 - LIHC	**50**	**3,88**	**3,88**	**0,50**	**2,86**	**4,99**	**3,27**	**3,55**	**1,40**	**-0,91**	**5,72**	**0,005**
30.0 - UCEC	7	4,65	4,71	0,74	3,50	5,65	2,78	2,60	1,39	1,22	5,70	0,013
26.0 - STAD	33	3,06	3,25	1,40	-0,07	5,89	3,91	3,89	1,30	1,04	7,95	0,014
4.0 - CHOL	9	3,66	3,71	0,54	2,74	4,38	2,41	2,31	1,40	0,85	5,94	0,032
21.0 - PCPG	3	1,28	1,20	0,16	1,15	1,50	4,14	3,65	1,78	2,25	6,53	0,086
20.0 - PAAD	4	2,84	2,49	0,99	1,97	4,39	4,13	3,82	0,67	3,62	5,28	0,108
5.0 - COAD	26	4,14	4,09	0,51	3,31	4,88	3,70	3,78	1,30	0,98	5,72	0,120
7.0 - ESCA	13	2,46	2,84	1,67	0,06	5,04	3,33	3,65	1,08	1,40	4,88	0,141
3.0 - CESC	3	4,84	4,85	0,43	4,31	5,36	2,49	3,79	1,98	-0,30	3,99	0,177
12.0 - KIRP	32	2,78	3,38	1,22	-0,27	4,17	3,12	3,09	1,10	-0,12	5,06	0,267
23.0 - READ	6	4,49	4,53	0,44	3,69	5,07	3,88	3,95	1,24	1,56	5,51	0,321
17.0 - LUSC	50	2,43	2,39	0,55	1,17	3,70	2,21	2,59	1,74	-1,64	6,59	0,405
29.0 - THYM	2	2,46	2,46	1,13	1,33	3,59	-0,10	-0,10	0,03	-0,13	-0,07	
24.0 - SARC	2	3,94	3,94	0,19	3,75	4,13	4,75	4,75	1,76	3,00	6,51	
8.0 - GBM	0											
14.0 - LGG	0											
19.0 - OV	0											
25.0 - SKCM	0											
27.0 - TGCT	0											

Os gráficos do tipo Boxplot fornecem uma visão mais sistêmica da distribuição dos dados. Na Figura 4 é possível visualizar a variação média e os outliers da expressão do gene *NT5E* para as duas combinações, tipo tumoral e não tumoral, para os tipos tumorais que tiveram uma significância do valor de $p \leq 0.01$.

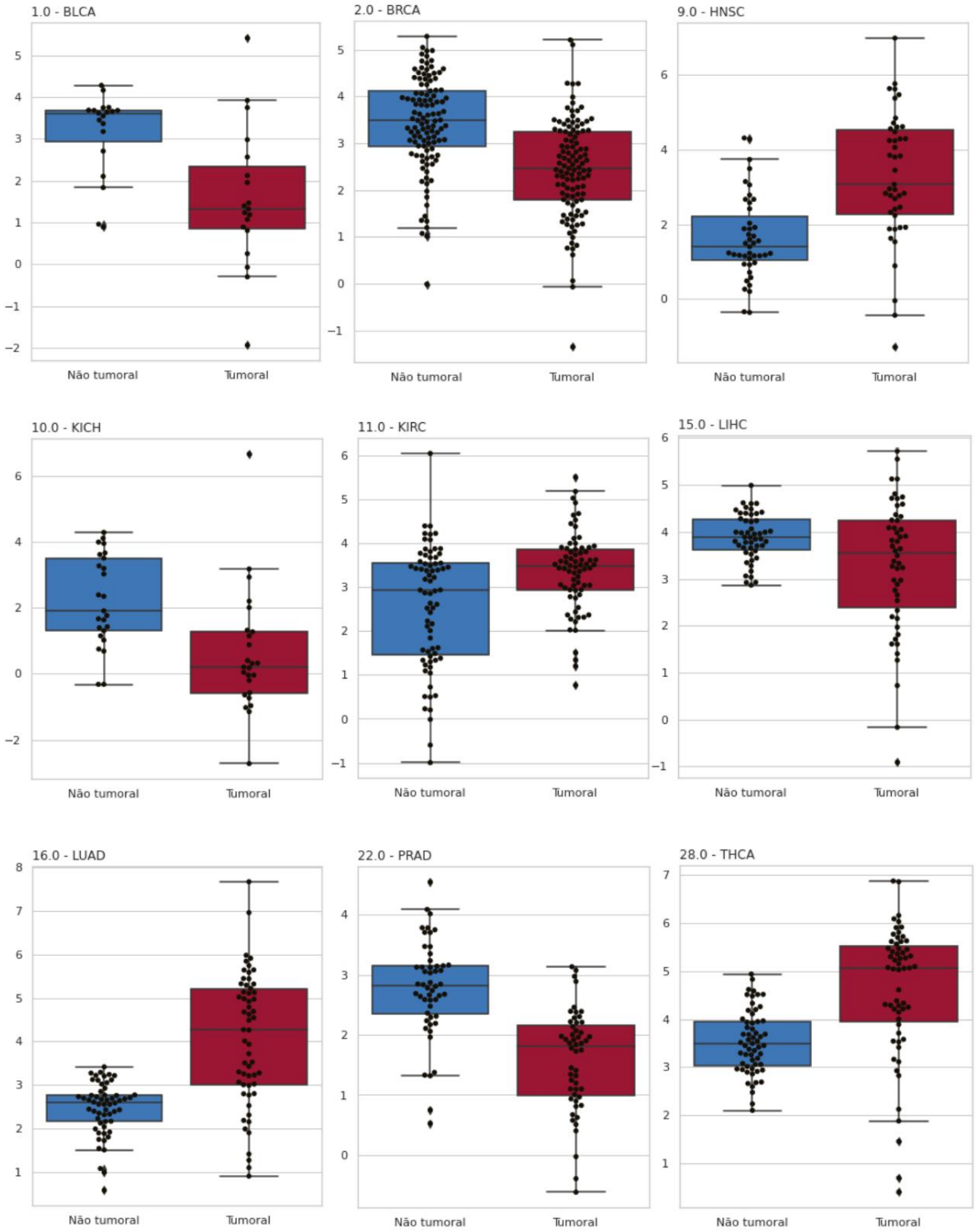

Figura 4: Gráficos do tipo Boxplot demonstrando a variação da expressão do gene *NT5E* para cada tipo tumoral com valor de $p \leq 0.01$. Fonte: Do autor.

Após a seleção dos tipos tumorais considerando a significância do gene *NT5E* entre tecido normal e tumoral, buscou-se classificar as amostras tumorais e normais através do modelo de aprendizado de máquina Florestas Aleatórias, considerando os níveis de metilação dos sítios do gene *NT5E*.

Como parâmetros, foram aplicados no modelo, somente o valor de expressão do gene *NT5E* e posteriormente foram combinados o valor de expressão do gene *NT5E* com cada sítio de metilação. Para a aplicação das amostras, foi necessário realizar uma seleção entre os dados dos sítios de metilação, pois em alguns sítios existiam valores faltantes, o que prejudicaria a análise, desta forma, sempre que existisse um valor de sítio faltante, a amostra era desconsiderada.

O tipo tumoral KICH (10.0) foi desconsiderado na análise do modelo pois o mesmo ficou desbalanceado, ou seja, não foram selecionadas amostras do tipo normal para este tipo tumoral (Tabela 6).

Tabela 6: Número de amostras utilizadas no modelo de aprendizado de máquina que teve como resultado os mapas de calor demonstrados abaixo, contendo o número separado de amostras normais e amostras tumorais de cada tipo tumoral analisado, tanto para a análise utilizando somente o valor da expressão do gene quanto para a combinação com cada sítio de metilação, por fim, também é demonstrado em vermelho o tipo tumoral desbalanceado. Fonte: Do autor.

	Expressão de gene		**Expressão de gene + metilações**	
Tipos tumorais	Normal	Tumoral	Normal	Tumoral
2.0 - BRCA	112	112	73	85
11.0 - KIRC	72	72	23	69
28.0 - THCA	59	59	50	59
16.0 - LUAD	58	58	19	48
22.0 - PRAD	52	52	35	52
15.0 - LIHC	50	50	41	50
9.0 - HNSC	43	43	20	43
10.0 - KICH	25	25	0	24
1.0 - BLCA	19	19	17	19

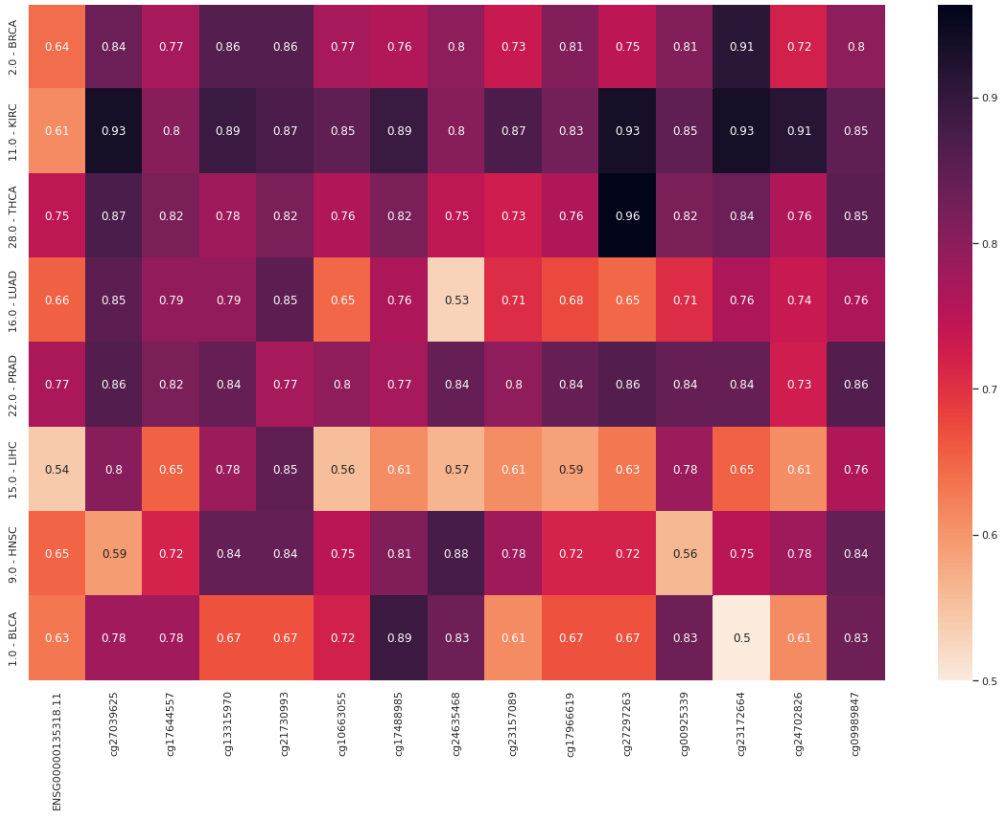

Figura 5: Mapa de calor dos resultados da Acurácia. Fonte: Do autor.

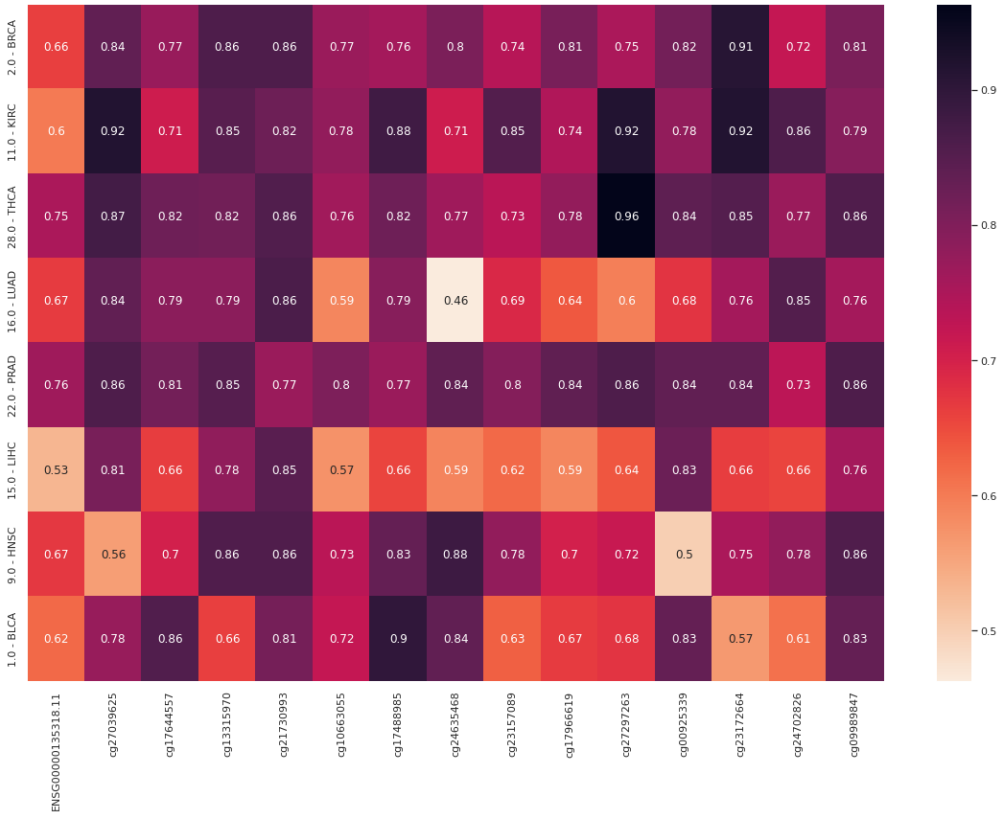

Figura 6: Mapa de calor dos resultados da Precisão. Fonte: Do autor.

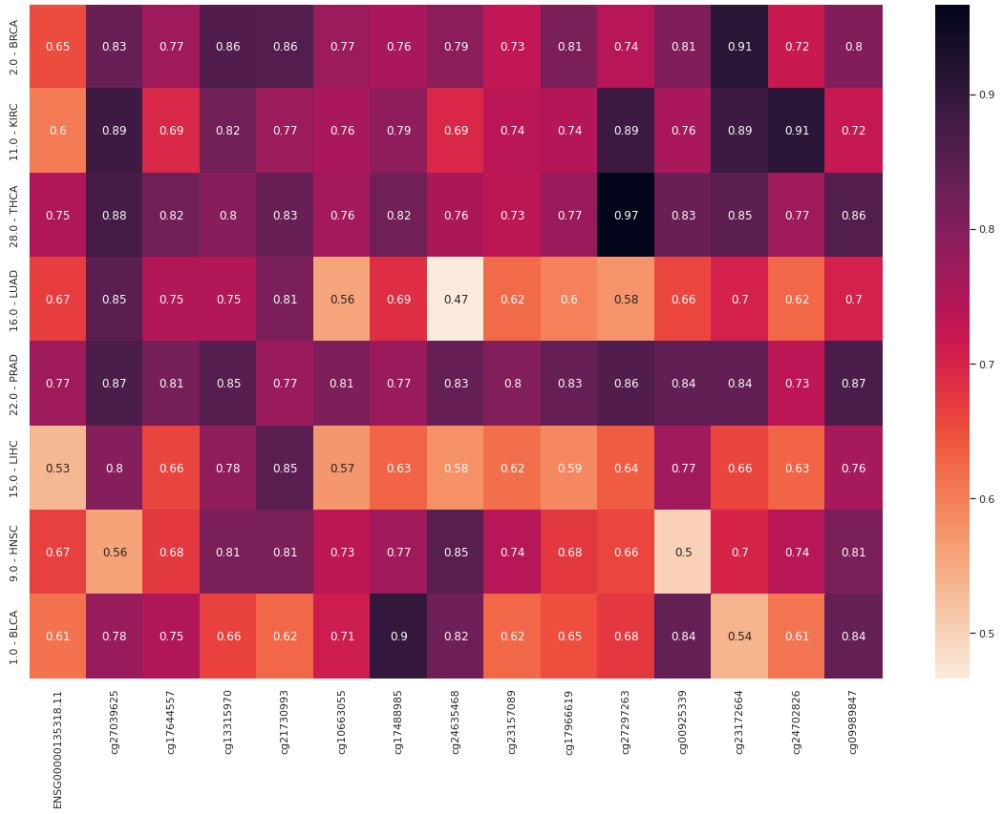

Figura 7: Mapa de calor dos resultados do Recall. Fonte: Do autor.

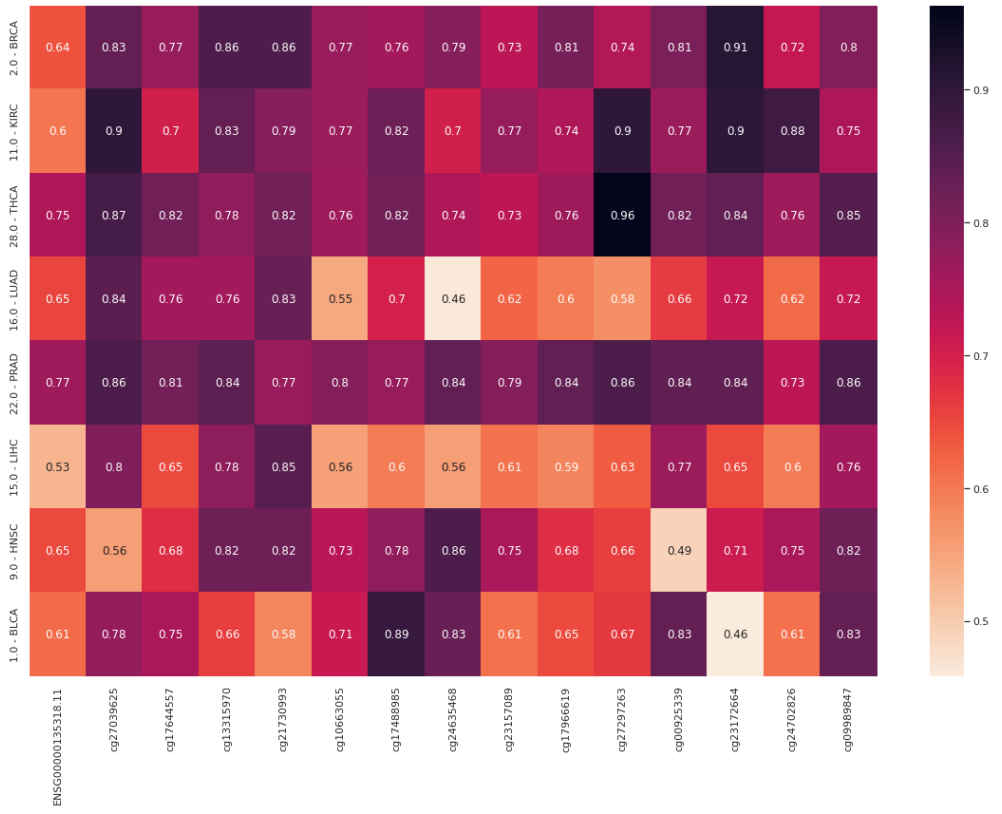

Figura 8: Mapa de calor dos resultados do F1-Score. Fonte: Do autor.

Para avaliação da eficiência do método aplicado, conforme apresentado nas Figuras 6-9, foram gerados os mapas de calor com os valores de quatro diferentes métricas: Acurácia, Precisão, Recall e F1-Score.

5 DISCUSSÃO

Considerando todo o processo realizado desde a importação dos dados até a conclusão de fato dos modelos preditivos, é possível perceber que no início do artigo, era apresentado um valor significativo de amostras, e ao longo de todas as seleções e etapas realizadas, resultou em um número ainda significativo, mas em proporções bem menores.

Foi necessário estabelecer filtros e analisar somente os dados pareados do gene *NT5E* e de tipos tumorais que se mostram com grande significância probabilística para chegar a um resultado satisfatório.

Entre os 33 tipos tumorais inicialmente importados, foi considerado aplicar ao modelo de aprendizado de máquina somente em 8 tipos tumorais, pois foi através da significância do valor de p, que foi possível estabelecer quais destes grupos de tipos tumorais teriam maiores chances de serem classificados a partir das caraterísticas do gene *NT5E*.

Também podemos destacar, que mesmo considerando utilizar as amostras com identificador do tipo 05, correspondente ao Novo Primário, o mesmo não entrou nas amostras finais pois não foi apresentado amostras pareadas do tipo 05 e 11, ficando com um grupo de amostras finais correspondentes apenas as amostras identificadas como 01 e 11, sendo respectivos ao Tumor Primário e Tecido Sólido Normal.

Após a seleção final dos dados por tipos tumorais e pelos identificadores de cada amostra respectivamente definidos como Tumor Primário (01) e Tecido Sólido Normal (11), foram aplicados as amostras ao modelo de aprendizado de máquina de duas formas diferentes, sendo a primeira forma utilizando somente o valor de expressão do gene e posteriormente aplicando novamente o valor de expressão do gene mas em conjunto com cada sítio de metilação, onde foi obtido como resultado final e mais significativa para o modelo, ou seja, os valores preditivos de F1-Score, representados na Figura 9.

Notou-se um certo desbalanceamento em alguns tipos tumorais, quando aplicado em conjunto com a expressão do gene e os sítios de metilação, pois em algumas amostras foram identificados dados faltantes, onde foi necessário desconsiderar a mesma e manter somente o seu par correspondente, caso ele tivesse algum valor. Um exemplo de desbalanceamento apresentado na Tabela 6, foi o tipo tumoral KIRC, LUAD, PRAD e HNSC, mas ainda apresentaram bons resultados de F1-Score, apresentado na Figura 9. Na Tabela 7 resume a comparação do F1-Score entre os valores gerados a partir da expressão do gene e dos valores menores e maiores da combinação entre a expressão do gene com os sítios de metilação para cada tipo tumoral.

Tabela 7: Comparação do F1-Score para eficiência da expressão do gene *NT5E* e sua combinação com seus sítios de metilação em diferentes tipos tumorais. Fonte: Do autor.

F1-Score					
	Expressão de gene	**Expressão de gene + metilações**			
Tipos tumorais	**Valor**	**Sítio**	**Valor menor**	**Sítio**	**Valor maior**
2.0 - BRCA	0.64	cg23157089	0.73	cg23172664	0.91
11.0 - KIRC	0.6	cg17644557, cg24635468	0.7	cg27039625, cg27297263, cg23172664	0.9
28.0 - THCA	0.75	cg23157089	0.73	cg27297263	0.96
16.0 - LUAD	0.65	cg24635468	0.46	cg27039625	0.84
22.0 - PRAD	0.77	cg24702826	0.73	cg27039625, cg27297263, cg09989847	0.86
15.0 - LIHC	0.53	cg10663055, cg24635468	0.56	cg21730993	0.85
9.0 - HNSC	0.65	cg00925339	0.49	cg24635468	0.86
1.0 - BLCA	0.61	cg23172664	0.46	cg17488985	0.89

Desta forma, nota-se que em todos os tipos tumorais aplicados no aprendizado de máquina, os valores de expressão do gene em combinação com cada tipo tumoral, tiveram melhores resultados, em relação a aplicação da expressão do gene de forma separada. Neste caso, sendo ainda mais específico, é

possível notar que em cada tipo tumoral tem um sítio de metilação com valor de predição mais predominante, ou seja, determinados sítios de metilação, geraram um valor superior ao demais.

6 CONCLUSÃO

Para o diagnóstico mais preciso de tumores e predição de progressão para um melhor ou pior prognóstico, mesmo com muitas ferramentas e tecnologias empregadas no mercado, ainda é um desafio, devido ao alto nível de complexibilidade de cada tipo tumoral.

Considerando que a identificação da progressão dos tumores e o conhecimento das especificidades de cada microambiente tumoral é muito importante para a obtenção de futuros resultados e procedimentos (Zheng and Xu, 2020), a geração de ferramentas que proporcionem aos profissionais auxílio preditivos com a utilização de inteligência artificial tem se mostrado muito importante e promissores. Ao analisar os dados genômicos presentes em plataforma públicas, pretende-se obter novos insights e gerar ferramentas que podem auxiliam em uma eventual decisão mais assertiva.

O software em questão, possibilitou a realização de coleta de dados de pacientes que se encontram em bases de dados do The Cancer Genome Atlas (TCGA) em um formato ainda pouco trabalhado e estruturado, ou seja, são diversos dados que estão disponíveis a qualquer usuário, mas que a obtenção e o filtro dos mesmos ainda necessitam de pessoas com conhecimentos técnicos para o seu manuseio. Esta ferramenta desenvolvida proporcionará que pesquisadores sem conhecimentos em tecnologias consigam obter os dados, filtrá-los e analisá-los do mesmo modo como foram realizadas ao longo deste artigo.

A expressão do gene *NT5E* juntamente com os sítios de metilação foi capaz de gerar um modelo adequado para auxiliar a caracterização dos tipos tumorais BRCA, KIRC, THCA, LUAD, PRAD, LIHC, HNSC e BLCA. Futuras análises em amostras frescas destes tipos tumorais são necessários para confirmar estas análises *in silico*. Ainda, como perspectivas, pretende-se analisar separadamente as amostras que foram classificadas corretamente ou não, tanto para o tipo normal como para o tipo tumoral, afim de conhecer melhor biologia e as especificidades de cada tipo tumoral.

7 REFERÊNCIAS

ALCEDO, Karel P.; BOWSER, Jessica L.; SNIDER, Natasha T. The elegant complexity of mammalian ecto-5′-nucleotidase (CD73). **Trends in Cell Biology**, v. 31, n. 10, p. 829–842, 2021.

ALLARD, Bertrand; ALLARD, David; BUISSERET, Laurence; et al. The adenosine pathway in immuno-oncology. **Nature Reviews Clinical Oncology**, v. 17, n. 10, p. 611–629, 2020.

ARAN, Dvir; CAMARDA, Roman; ODEGAARD, Justin; et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. **Nature Communications**, v. 8, n. 1, p. 1077, 2017.

CASALINO, Laura; VERDE, Pasquale. Multifaceted Roles of DNA Methylation in Neoplastic Transformation, from Tumor Suppressors to EMT and Metastasis. **Genes**, v. 11, n. 8, p. 922, 2020.

CHEN, Qiangda; PU, Ning; YIN, Hanlin; et al. CD73 acts as a prognostic biomarker and promotes progression and immune escape in pancreatic cancer. **Journal of cellular and molecular medicine**, v. 24, n. 15, p. 8674–8686, 2020.

COOPER, Lee Ad; DEMICCO, Elizabeth G.; SALTZ, Joel H.; et al. PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective. **The Journal of pathology**, v. 244, n. 5, p. 512–524, 2018.

IWENDI, Celestine; BASHIR, Ali Kashif; PESHKAR, Atharva; et al. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. **Frontiers in Public Health**, v. 8, p. 357, 2020.

JEONG, Young Ju; OH, Hoon Kyu; CHOI, Hye Ryeon; et al. Methylation of the NT5E Gene Is Associated with Poor Prognostic Factors in Breast Cancer. **Diagnostics**, v. 10, n. 11, p. 939, 2020.

KÄÄRIÄINEN, Anni; PESOLA, Vilma; DITTMANN, Annalena; et al. Machine learning identifies robust matrisome markers and regulatory mechanisms in cancer. **International journal of molecular sciences**, v. 21, n. 22, p. 8837, 2020.

KANANI, Pratik; PADOLE, Mamta. Deep learning to detect skin cancer using google colab. **International Journal of Engineering and Advanced Technology Regular Issue**, v. 8, n. 6, p. 2176–2183, 2019.

LI, Yuanyuan; KANG, Kai; KRAHN, Juno M.; et al. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. **BMC Genomics**, v. 18, n. 1, p. 508, 2017.

LO NIGRO, C.; MONTEVERDE, M.; LEE, S.; et al. NT5E CpG island methylation is a favourable breast cancer biomarker. **British Journal of Cancer**, v. 107, n. 1, p. 75–83, 2012.

RASCHKA, Sebastian; PATTERSON, Joshua; NOLET, Corey. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. **Information**, v. 11, n. 4, p. 193, 2020.

SANNI, Rachana R; GURUPRASAD, H. S. Analysis of performance metrics of heart failured patients using Python and machine learning algorithms. **Global Transitions Proceedings**, v. 2, n. 2, p. 233–237, 2021. (International Conference on Computing System and its Applications (ICCSA- 2021)).

SPAINHOUR, John CG; LIM, Hong Seo; YI, Soojin V; et al. Correlation Patterns Between DNA Methylation and Gene Expression in The Cancer Genome Atlas. **Cancer Informatics**, v. 18, p. 1176935119828776, 2019.

TYRALIS, Hristos; PAPACHARALAMPOUS, Georgia. Variable Selection in Time Series Forecasting Using Random Forests. **Algorithms**, v. 10, n. 4, p. 114, 2017.

WANG, H.; LEE, S.; NIGRO, C. Lo; et al. NT5E (CD73) is epigenetically regulated in malignant melanoma and associated with metastatic site specificity. **British Journal of Cancer**, v. 106, n. 8, p. 1446–1452, 2012.

WAY, Gregory P.; SANCHEZ-VEGA, Francisco; LA, Konnor; et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. **Cell Reports**, v. 23, n. 1, p. 172-180.e3, 2018.

ZHENG, Chunlei; XU, Rong. Predicting cancer origins with a DNA methylation-based deep neural network model. **PLOS ONE**, v. 15, n. 5, p. e0226461, 2020.