Guilherme Ferreira Lourenço

Análise de Linhagens Celulares de Melanoma via Aprendizado de Máquina

São José dos Campos - Brasil

Outubro de 2021

Guilherme Ferreira Lourenço

Análise de Linhagens Celulares de Melanoma via Aprendizado de Máquina

Relatório apresentado à Universidade Federal
de São Paulo referente à Iniciação Científica.

Aluno: Guilherme Ferreira Lourenço

Docente: Profa. Lilian Berton

Universidade Federal de São Paulo - UNIFESP

Instituto de Ciência e Tecnologia - Campus São José dos Campos

São José dos Campos - Brasil

Outubro de 2021

Resumo

Este projeto visa analisar características de redes de interações de proteínas expressas por diferentes linhagens celulares humanas de melanoma em busca de padrões dentro dos diferentes grupos e aplicar técnicas de aprendizado de máquina (AM) para o reconhecimento de melanoma. Ademais, visa oferecer uma oportunidade ao aluno de se aprofundar na área de AM, com uma experiência palpável em desenvolvimento de um projeto que envolva várias etapas – levantamento das features de dados, aplicação dos modelos, validação dos resultados, bem como a interpretação dos resultados. Além de obter experiência com a realização de pesquisa e trabalhos acadêmicos.

Palavras-chaves: cancer de pele, redes de proteínas, redes complexas, aprendizado de máquina.

Lista de ilustrações

Figura 1 – Demonstração dos graus dos vértices em uma rede simples, onde cada vértice acompanha seu valor de grau. Fonte: (1). 12
Figura 2 – Exemplo de transitividade tripla em uma rede direcionada, onde u se conecta a v e v se conecta a w, logo u se liga a w. Fonte: (1). 13
Figura 3 – Exemplo de uso da Árvore de Decisão para classificação de um mamão. Fonte: (2). 15
Figura 4 – Representação dos dados referentes às proteínas da linhagem celular maligna do melanoma de fonte primária. Fonte: Autor. 23
Figura 5 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um recém-nascido. Fonte: Autor. . . . 24
Figura 6 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um paciente com melanoma no pulmão. Fonte: Autor. 25
Figura 7 – Representação dos dados referentes às proteínas da linhagem celular de um local metastático do nodo pulmonar de um paciente que tinha melanoma. Fonte: Autor. 26
Figura 8 – Representação dos dados referentes às proteínas da linhagem celular derivada de derrame pleural de um paciente com melanoma metastático. Fonte: Autor. 27
Figura 9 – Representação dos dados referentes às proteínas da linhagem celular maligna do melanoma de fonte primária. Fonte: Autor. 28
Figura 10 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um recém-nascido. Fonte: Autor. . . . 29
Figura 11 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um paciente com melanoma no pulmão. Fonte: Autor. 30
Figura 12 – Representação dos dados referentes às proteínas da linhagem celular de um local metastático do nodo pulmonar de um paciente que tinha melanoma. Fonte: Autor. 31
Figura 13 – Representação dos dados referentes às proteínas da linhagem celular derivada de derrame pleural de um paciente com melanoma metastático. Fonte: Autor. 32
Figura 14 – Diagrama de Venn das proteínas de cada amostra dos dados. Fonte: Autor. 33
Figura 15 – Matriz de Confusão Multi-classe dos resultados do classificador K-Nearest Neighbours. Fonte: Autor. 34

Figura 16 – Relatório de Classificação dos resultados do classificador K-Nearest Neighbours. Fonte: Autor. 35
Figura 17 – Matriz de Confusão Multi-classe dos resultados do classificador Naive Bayes. Fonte: Autor. 36
Figura 18 – Relatório de Classificação dos resultados do classificador Naive Bayes. Fonte: Autor. 36
Figura 19 – Matriz de Confusão Multi-classe dos resultados do classificador Decision Tree. Fonte: Autor. 37
Figura 20 – Relatório de Classificação dos resultados do classificador Decision Tree. Fonte: Autor. 37
Figura 21 – Matriz de Confusão Multi-classe dos resultados do classificador Bagging Meta-Estimator (Decision Tree). Fonte: Autor. 38
Figura 22 – Relatório de Classificação dos resultados do classificador Bagging Meta-Estimator (Decision Tree). Fonte: Autor. 39
Figura 23 – Matriz de Confusão Multi-classe dos resultados do classificador Gradient Tree Boosting. Fonte: Autor. 39
Figura 24 – Relatório de Classificação dos resultados do classificador Gradient Tree Boosting. Fonte: Autor. 40
Figura 25 – Matriz de Confusão Multi-classe dos resultados do classificador Histogram-Based Gradient Boosting Classification Tree. Fonte: Autor. 41
Figura 26 – Relatório de Classificação dos resultados do classificador Histogram-Based Gradient Boosting Classification Tree. Fonte: Autor. 41
Figura 27 – Gráfico de barras apresentando as medidas de avaliação geral de todos os algoritmos em conjunto. Fonte: Autor. 42

Sumário

1 INTRODUÇÃO 7

2 OBJETIVOS 9
2.1 Objetivo geral 9
2.2 Objetivos Específicos 9

3 FUNDAMENTAÇÃO TEÓRICA 11
3.1 Câncer de pele 11
3.2 Redes complexas 11
3.3 Aprendizado de Máquina 13
3.3.1 Algoritmos de aprendizado supervisionado 14

4 MATERIAIS E MÉTODOS 19
4.1 Dataset 19
4.2 Bibliotecas 19
4.3 Fluxo do trabalho 20

5 RESULTADOS 23
5.1 Análise visual 23
5.2 Resultado dos classificadores 33
5.3 Resultado dos classificadores ensembles 38
5.4 Considerações finais 42

REFERÊNCIAS 43

1 Introdução

Segundo a Sociedade Brasileira de Dermatologia, o câncer de pele é causado pelo crescimento anormal e descontrolado das células que compõem a pele. Essas células se dispõem formando camadas e, de acordo com as que forem afetadas, são definidos os diferentes tipos de câncer. Os mais comuns são os carcinomas, porém, o melanoma é mais raro e letal que os carcinomas sendo o tipo mais agressivo de câncer da pele devido à sua alta possibilidade de provocar metástase (disseminação do câncer para outros órgãos).

O melanoma se trata de um tumor maligno que afeta as células produtoras de melanina (melanócitos) causando a multiplicação celular descontrolada, podendo ser identificado em qualquer parte do corpo, geralmente na forma de uma mancha, pinta ou sinais na pele. Segundo a World Cancer Research Fund[1], o melanoma é o 19º diagnóstico mais comum do mundo. Já no Brasil, este tipo de tumor representa cerca de 3% dos registros, sendo 8.450 novos casos apenas em 2020[2]. Embora não seja tão frequente em relação aos outros cânceres, esse é o mais fatal dos tumores malignos cutâneos, uma vez que facilmente se espalha para outros órgãos, podendo causar metástase no paciente. Ainda assim, o prognóstico é considerado bom quando diagnosticado nos primeiros estágios da doença. Portanto, estudar essa patologia profundamente é essencial para entender mais o seu comportamento e efeitos no corpo humano, para que se possa aperfeiçoar seu diagnóstico e tratamento.

As proteínas intracelulares podem ser representadas através de grafos, onde as conexões entre elas representam as interações proteína-proteína existente. Essas interações podem ser diretas, também chamadas interações físicas (por exemplo, complexos proteicos), ou indiretas, que também são nomeadas de interações funcionais (quando as proteínas participam do mesmo processo biológico) (3). O tipo de interação a ser representada depende de qual será a amostra e qual será o objetivo da construção da rede.

Outra técnica computacional que auxilia a análise de padrões em dados é o Aprendizado de Máquina, o qual pode ser definido como uma área da Inteligência Artificial (IA) que foca em construir programas capazes de identificar padrões, efetuar predições e melhorar desempenho com base em amostras de dados previamente expostos (4). Dentro deste campo da IA, existem divisões de aprendizado, tais como aprendizado supervisionado e aprendizado não supervisionado. Atualmente, técnicas computacionais têm obtido sucesso no auxílio ao diagnóstico de diversas doenças, nesse trabalho será explorado técnicas de redes complexas e aprendizado de máquina no reconhecimento de melanoma.

[1] https://www.wcrf.org/dietandcancer/cancer-trends/skin-cancer-statistics
[2] https://www.inca.gov.br/tipos-de-cancer/cancer-de-pele-melanoma

2 Objetivos

2.1 Objetivo geral

Analisar características topológicas de redes de interações de proteínas expressas por diferentes linhagens celulares humanas de melanoma em busca de padrões característicos dos diferentes grupos. Empregar essas características em algoritmos de Aprendizado de Máquina (AM) para auxílio na predição de melanoma.

2.2 Objetivos Específicos

- Aplicar e analisar medidas de redes complexas para caracterização dos dados.
- Comparar o desempenho entre algoritmos tradicionais de classificação em AM.
- Atribuir significado biológico para as análises.
- Participar das oficinas e reuniões do programa WASH.

3 Fundamentação Teórica

3.1 Câncer de pele

Cummins *et al.* (2006) define o melanoma como um tumor maligno que tem sua ação ocorrendo em quaisquer tecidos que contenham melanócitos (5), ou seja, se refere a uma malignidade que afeta as células produtoras de melanina, podendo prejudicar partes do corpo humano, desde a pele até lugares não cutâneos, como o tecido que reveste a cavidade bucal, na íris dos olhos, pulmões e até o sistema nervoso central, embora seja mais comum ocorrer na pele.

O melanoma é uma categoria de câncer de pele que ocorre cerca de 232.100 novos casos por ano no mundo todo, dos quais a maioria se situam nas Américas e na Oceania. Apesar de corresponder a uma incidência relativamente baixa em relação aos outras categorias de malignidades cutâneas, é o tipo com maior taxa de letalidade, causando cerca de 55.500 mortes anualmente (6).

Grandes motivos que elevam chances de uma ocorrência incluem demasiada exposição aos raios ultravioletas, histórico na família (herança de uma mutação genética) e presença de sardas no indivíduo. Devido a isto, é estimado que uma redução de 10% da camada de ozônio, pode resultar em mais de 4.500 casos de melanoma (7), dado que a exposição aos raios UV presentes na luz solar é um fator significativo que leva ao desenvolvimento do melanoma. Logo, o aumento no número de casos pode estar ligado ao enfraquecimento da proteção da camada de ozônio, devido a sua função de agir como filtro da radiação UV.

3.2 Redes complexas

Uma rede é definida de forma simples como um conjunto de pontos interligados por linhas (8), ou seja, um grafo contendo vértices (também chamados de Nós) e arestas que os interligam. Esta estrutura abstrata é muito utilizada para representar relações entre múltiplos dados independentes da área, portanto é uma poderosa estrutura das áreas de ciência da computação, biologia, ciências humanas e várias outras que possibilitam o uso de redes. O que diferencia uma rede complexa de um grafo simples é a presença de algumas propriedades topográficas que podem ser medidas para várias análises como a importância que cada vértice tem no sistema da rede. Para o trabalho em questão, foram utilizadas as seguintes métricas de uma rede complexa:

Degree

Esta medida, que também pode ser chamada em português como Grau, G_i, se refere ao número de arestas a_{ij} ligadas em um dado vértice i. No caso da rede com arestas não-dirigidas, esta métrica pode ser definida por:

$$G_i = \sum_{j \in N} a_{ij}$$

Este é um método para medir a centralidade de um nó e, em muitos casos, é um valor relevante que pode indicar a importância que o vértice tem no funcionamento do sistema da rede (1). Um exemplo de uma rede com diferentes graus é dado na Figura 1.

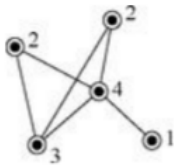

Figura 1 – Demonstração dos graus dos vértices em uma rede simples, onde cada vértice acompanha seu valor de grau. Fonte: (1).

Closeness

Closeness é outra métrica de centralidade utilizada no trabalho em questão. A medida é analisada pela média das distâncias geodésicas d_{ij} (menor caminho de arestas entre dois vértices i e j) entre um vértice e os demais nós da rede (1), sendo este calculado por:

$$C_i = \frac{n}{\sum_j d_{ij}}$$

Embora seja uma métrica que, em alguns casos, pode apresentar dificuldade para diferenciar alguns vértices mais centrais de outros menos centrais devido ao range de valores que os resultados possuem, é uma medida muito utilizada em redes complexas variadas (1).

Clustering

Esta métrica, quando aplicada em um vértice i, analisa a transitividade entre ele e seus vizinhos, ou seja, verifica a probabilidade de que dois nós vizinhos aleatórios j e m sejam também, vizinhos entre si, formando assim um "triângulo" entre estes três vértices (8). Um exemplo é dado na Figura 2.

Essa análise possibilita verificar numericamente a chance de existir "buracos estruturais", o que pode significar que o vértice em análise tem uma grande importância e, neste sentido, a métrica se torna uma medida de centralidade também (1). Com isso, é possível definir o valor de Clustering C_i de um dado vértice i da seguinte forma:

$$C_i = \frac{(2 \times l_i)}{k_i(k_i - 1)}$$

Nota-se que a definição utiliza k_i para representar o conceito de grau do vértice i para obter o valor do número total de pares de seus vizinhos e, utilizando l_i, tem-se o número de pares de vizinhos conectados ao vértice i e estão conectados entre si. Dessa forma, obtém-se a probabilidade de haver transitividade entre um vértice e um par de vizinhos.

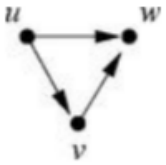

Figura 2 – Exemplo de transitividade tripla em uma rede direcionada, onde u se conecta a v e v se conecta a w, logo u se liga a w. Fonte: (1).

Betweenness

Esta medida de centralidade utiliza do conceito de menor caminho geodésico entre dois vértices, visto que, para definir B_i, a métrica observa as menores distâncias entre dois vértices j e k, e que passam pelo vértice i, obtendo o valor de d_{jk} (i) (8). A definição para obter o valor para cada vértice se dá pela somatória das razões entre $d_{jk}(i)$ e d_{jk} (1):

$$B_i = \sum_{j,k \in G, j \neq k} \frac{d_{jk}(i)}{d_{jk}}$$

Dessa forma, pode-se considerar que quanto maior o valor, maior influência que o vértice tem no sistema da rede.

3.3 Aprendizado de Máquina

O desenvolvimento tecnológico na área da computação gerou grandes impactos na humanidade e, com o passar dos anos, a complexidade das aplicações de conceitos computacionais foi evoluindo cada vez mais. Com isso, surgiu a hipótese de que computadores poderiam conseguir aprender tarefas. Logo, criou-se o campo de Aprendizado de Máquina

que, segundo (9), empenha-se em responder a pergunta de "como construir programas computacionais que aprimoram automaticamente com experiência".

Existem diferentes categorias de aprendizagem que definem subcampos da ciência do Aprendizado de Máquina e o emprego de cada um deles depende dos dados do problema. Um desses segmentos é o supervisionado que, de acordo com (10), assume que os dados contenham uma informação "alvo" que ajudará na classificação de novas instâncias. Essa informação é denominada por (11) como "labels", que seriam as diferentes categorias do problema em que cada instância de dado pode ser classificado. Já os algoritmos de aprendizado não supervisionado, como (11) define, não são utilizadas as tais "labels" para supervisionar como os dados devem ser individualmente classificados. Assim, os dados podem ser agrupados por meio de medidas de similaridade. Outra classe de Aprendizado de Máquina é o aprendizado semissupervisionado, (12) afirma que esse método compreende técnicas que combinam dados com "label" e dados sem "label" para treinamento do algoritmo. Dessa forma, o conjunto de dados é nomeado dados parcialmente rotulados.

3.3.1 Algoritmos de aprendizado supervisionado

K-Nearest Neighbours

De acordo com (9), este algoritmo utiliza o conceito da distância euclidiana para relacionar dados entre as classes do problema. Para isto, cada dado X é descrito como um vetor de características ($a_1(X)$, $a_2(X)$, ..., $a_n(X)$), onde cada $a_r(X)$ representa um atributo do dado.

Dessa forma, para classificar um dado X_i, deve-se calcular cada distância Euclidiana entre ele e os dados vizinhos, para isso o espaço entre duas instâncias X_i e X_j pode ser definida por d(X_i, X_j) ou $\|X_i - X_j\|$, *onde* :

$$d(X_i, X_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

Obtendo os valores de todas as distâncias entre X_i e seus "vizinhos", o algoritmo compara as classes dos K dados mais próximos e retorna a label mais frequente entre eles, concluindo que X_i provavelmente pertence a esta categoria. Note que K deve ser um valor inteiro previamente determinado.

Naive Bayes

No caso do algoritmo de Naive Bayes, o método calcula a probabilidade de cada classe utilizando o teorema de Bayes com a premissa de que os atributos são independentes entre si (2). Ao obter o valor de cada probabilidade, o algoritmo define a classificação do dado de entrada segundo a classe provável. Para isso, o algoritmo utiliza da seguinte

fórmula para classificação de um dado vetor de atributos X = (X_1, ..., X_d) em alguma "label" y, sendo que y pertence a um conjunto finito Y de possíveis classes:

$$z = argmax_{y \in Y} P(Y = y) \prod_{i=1}^{d} P(X_i = x_i | Y = y)$$

Após estimar a probabilidade de cada classe com base na frequência de cada atributo, o algoritmo é capaz de relacionar novas instâncias com alguma classificação existente no conjunto finito Y (9).

Decision Tree

Uma árvore de decisão é uma estratégia vastamente utilizada em vários algoritmos para predizer a "label" de uma instância x, assim como os outros algoritmos. A diferença é que esse algoritmo classifica de forma booleana com uma função representada por uma árvore em que se vai do nó raiz até os nós-folha (9).

No exemplo da Figura 3 é possível notar o uso da Árvore de Decisão para demonstração do algoritmo de forma simples, em que cada nó interno representa um atributo do mamão a ser classificado, como cor e maciez, as decisões são feitas de forma booleana segundo as características da fruta e os nós-folha são as classificações finais, que no caso são se o mamão é saboroso ou não.

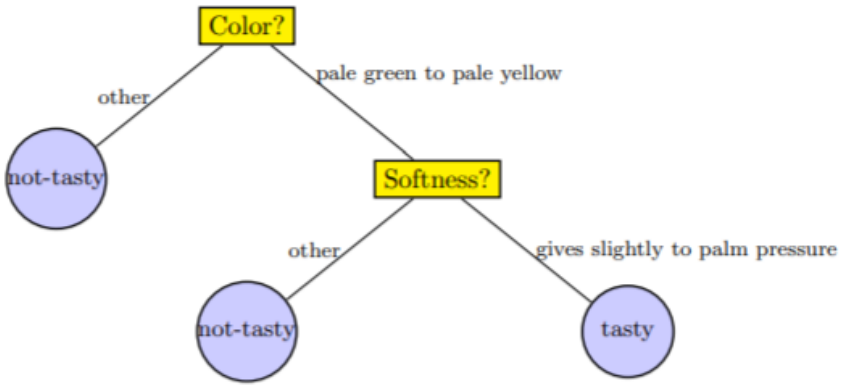

Figura 3 – Exemplo de uso da Árvore de Decisão para classificação de um mamão. Fonte: (2).

Gradient Tree Boosting

Este algoritmo se diferencia dos anteriores, pois utiliza um método diferente que melhora a robustez dos resultados, este método é chamado "Ensemble". Os algoritmos que adotam esta técnica trabalham a combinação de outros classificadores para gerar um resultado com menor possibilidade de erros (13).

Há várias maneiras de combinar os classificadores, mas no caso de Gradient Tree Boosting a técnica utilizada é a de Impulso (Boosting), que consiste em combinar várias Árvores de Decisão de forma serial, assim utilizando o resultado de uma iteração como um *bias* para minimizar o erro da Árvore da próxima iteração (14), em outros termos, a Árvore seguinte tentará corrigir o erro da Árvore anterior, assim gerando um algoritmo único muito mais robusto que apenas uma Árvore de Decisão. Dessa forma, o algoritmo trabalha para predizer uma classificação, focando em minimizar o erro em suas iterações (15).

Histogram-Based Gradient Boosting Classification Tree

Este algoritmo se trata de uma variante da Gradient Tree Boosting baseada na estratégia de agrupar os dados em valores inteiros (geralmente 256 grupos), de forma que acabe reduzindo as divisões nas árvores. Assim, para cada nó gerado na árvore, o algoritmo itera sobre as instâncias obtendo alguns valores estatísticos que preenchem um histograma para aquela "feature", o que permite o algoritmo utilizar números inteiros para as predições de melhores partições nas árvores, ao invés de valores contínuos que necessitam ser ordenados (14).

Bagging meta-estimator

No caso de Bagging meta-estimator, a técnica de "Ensemble" utilizada é a de Média (Averaging), ou seja, são utilizados alguns classificadores indepentendemente e depois estes resultados individuais são incorporados em uma única solução (13).

Logo, o algoritmo também tem como entrada algum classificador (por exemplo, KNN ou Árvore de Decisão) que será utilizado em partes aleatórias das instâncias de teste para a construção da solução final, este classificador é denominado estimador de base. Esta estratégia é utilizada para minimizar a variância do estimador base, tornando-o mais consistente e preciso.

Medidas de avaliação

Para todos os algoritmos, foi utilizado o método de teste K-fold Cross Validation, que consiste na divisão dos dados em K partições aleatórias de tamanhos aproximadamente iguais e, então, depois são feitos treinamentos do algoritmo para cada K partição diferente, dessa forma cada instância será utilizada no conjunto de treinamento e no conjunto de teste do algoritmo (16).

Para a avaliação dos resultados, foram utilizados alguns conceitos da estatística, como a Matriz de Confusão, que se trata de um meio muito comum para avaliar o erro de classificações, visto que, de maneira simples, demonstra quatro contagens, informações

importantes para inspecionar a classificação de cada classe (4). São essas:

- Verdadeiro Positivo (VP): O dado pertence à classe em questão e foi classificado corretamente;
- Falso Positivo (FP): O dado não pertence à classe em questão, mas o modelo previu que pertence (Erro do Tipo 1);
- Falso Negativo (FN): O dado pertence à classe em questão, mas o modelo previu que não pertence (Erro do Tipo 2);
- Verdadeiro Negativo (VN): O dado não pertence à classe em questão e foi classificado corretamente.

Dessas informações adquiridas pela Matriz de Confusão, é possível calcular outras métricas muito importantes para avaliar o modelo utilizado na classificação dos dados. Dentre elas, tem-se a acurácia, uma medida de avaliação vastamente utilizada que calcula a taxa de acertos do modelo (4). Para definir o valor da acurácia, podem ser utilizados os números da Matriz de Confusão da seguinte forma:

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

Assim, Acc representa a porcentagem de acertos do modelo.

Outro exemplo de medida de avaliação utilizada é a Precisão, em que se obtém a taxa de acertos positivos dentre todas as classificações como positivo (4), ou seja, reponde a pergunta "entre todos os dados classificados como positivo, quantos foram relacionados corretamente?". Para calcular o valor, tem-se a seguinte definição:

$$Prec = \frac{VP}{VP + FP}$$

Normalmente, essa medida é seguida de outra denominada Revocação (também conhecida como Sensibilidade), que ajuda a inferir sobre a capacidade do modelo de classificar uma classe (4), ou seja, responde a pergunta "quando a classe X é a label real do dado, qual a frequência que o modelo prediz que o dado pertence a essa classe X?". A medida pode ser definida da seguinte maneira:

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

Por fim, a medida F1-score calcula a média harmônica entre Precisão e Revocação e retorna um valor ente 0 e 1, que reflete a qualidade geral do modelo classificador. O valor desta métrica pode ser obtido por:

$$F = \frac{2}{\frac{1}{Rev} + \frac{1}{Prec}}$$

4 Materiais e métodos

4.1 Dataset

O material de estudo é composto por triplicatas de proteínas identificadas a partir da digestão tripsínica de proteínas intracelulares provenientes de lisados celulares adquiridas da American Type Culture Collection. Foram utilizados dados de espectrometria de massas de alta resolução processados pelo SW MaxQuant. As proteínas em questão foram divididas nos cinco grupos a seguir:

- HS68: linhagem celular de fibroblastos da pele saudável de um recém-nascido;
- HS895SK: linhagem celular de fibroblastos da pele saudável de um paciente com melanoma no pulmão;
- A375: linhagem celular maligna de melanoma de fonte primária;
- HS985T: linhagem celular obtida a partir de um local metastático do nodo pulmonar de um doente que tinha melanoma;
- SH4: linhagem celular derivada de derrame pleural de um paciente com melanoma metastático.

Foram utilizados os dados de label-free quantification (LFQ intensity) do arquivo de grupos de proteínas de saída do MaxQuant. Estes dados podem ser entendidos como valores de expressão proteica para as linhagens de estudo. Foram removidas proteínas onde havia a ausência de dados nas replicatas e coeficiente de variação maior que 20% entre os valores de LFQ nas triplicatas biológicas.

As interações entre as proteínas foram obtidas através de dados do repositório online público com interações proteicas catalogadas, o STRING database, mantendo as conexões em que ambas as proteínas foram identificadas como presentes nas linhagens.

4.2 Bibliotecas

Para efetuar os processos de estudo do trabalho em questão, foram utilizadas algumas bibliotecas implementadas em código Python.

Pandas

Para a manipulação dos dados, foi utilizada a biblioteca Pandas. Este recurso fornece funções e métodos capazes de obter e organizar dados em um DataFrame, que

se trata de um objeto com indexação integrada que facilita o manuseio de dados. Tais funcionalidades foram utilizadas para organização e filtragem de amostras, até fusão de algumas tabelas (17).

Numpy

Fundamental para a computação científica, a biblioteca Numpy foi utilizada para auxílio de operações que necessitam de ferramentas matemáticas mais completas e robustas. O recurso conta com várias soluções que facilitam operações matemáticas com arrays e outros operadores. Para o estudo, numpy foi utilizado no auxílio da geração de visualizações gráficas dos resultados (18).

Matplotlib

Possibilitando a melhoria da visualização dos dados e dos resultados de maneira que torne possível uma análise mais precisa, foi utilizada a biblioteca Matplotlib, um recurso muito conhecido para geração de diferentes ilustrações em Python (19). Esta ferramenta foi fundamental para obtenção de gráficos e outros materiais ilustrativos que demonstram o comportamento da amostra e dos resultados de cada modelo testado.

Scikit Learn

Por fim, a biblioteca Scikit Learn, a mais utilizada por se tratar uma solução open-source que contribui com a disponibilização de vários classificadores de aprendizado de máquina, contando também com recursos de abordagem para testes, como K-Fold Cross Validation, até métricas de avaliação dos modelos, como Matriz de Confusão de múltiplas classes e relatórios de classificação (20). Basicamente, esta ferramenta possibilita o uso dos modelos de aprendizado de máquina para treinamento e teste com os dados disponibilizados e, então, gera resultados que podem ser visualizados com auxílio da biblioteca Matplotlib.

4.3 Fluxo do trabalho

O estudo se desenvolve por meio do uso de redes complexas, área que estuda uma categoria de grafo não trivial com propriedades topológicas. Para analisar essas redes e avaliar a importância dos vértices, existem várias medidas para auxiliar no estudo dos dados em questão. Dentre elas, serão utilizadas: Degree, Clustering, Closeness e Betweenness.

O estudo também visa utilizar técnicas de AM para compreender e identificar padrões dos dados das proteínas. Essas técnicas consistem no desenvolvimento de algoritmos que melhoram automaticamente sua eficácia através de experiência (9). O segmento de AM a ser aplicado é o Aprendizado Supervisionado, que se resume na construção de um algoritmo capaz de associar um conjunto de dados $\{x_0, x_1, ..., x_n\}$ a um conjunto de resultados $\{y_0, y_1, ..., y_n\}$. Exemplos de modelos que serão utilizados: K-Nearest Neighbours,

Árvore de Decisão, Naive Bayes e Support Vector Machine, Redes Neurais.

5 Resultados

5.1 Análise visual

A análise visual consiste na observação crítica de gráficos que representam os dados de forma que seja possível inferir algumas características sobre o conjunto todo de dados das linhagens celulares. Portanto, após normalizar os valores das medidas de centralidade, foram gerados os seguintes gráficos:

Boxplots

- Fonte primária do melanoma (A375):

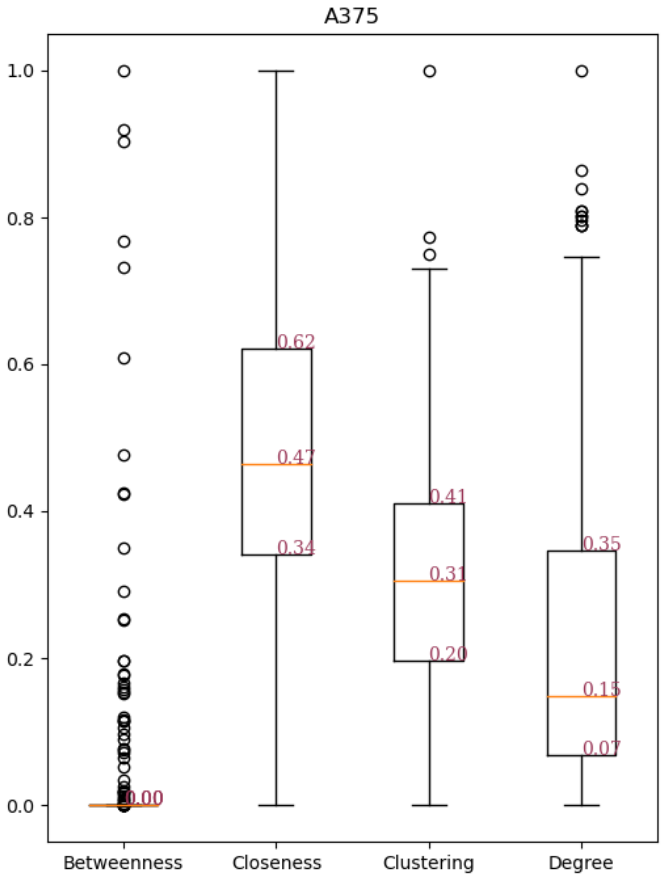

Figura 4 – Representação dos dados referentes às proteínas da linhagem celular maligna do melanoma de fonte primária. Fonte: Autor.

Nota-se na figura 4 que entre as quatro medidas de centralidade, Betweenness chama atenção pelo seu grande número de outliers quando comparada às outras métricas, além de ter um valor de média que aproxima de zero.

- Pele saudável de recém-nascido (HS68):

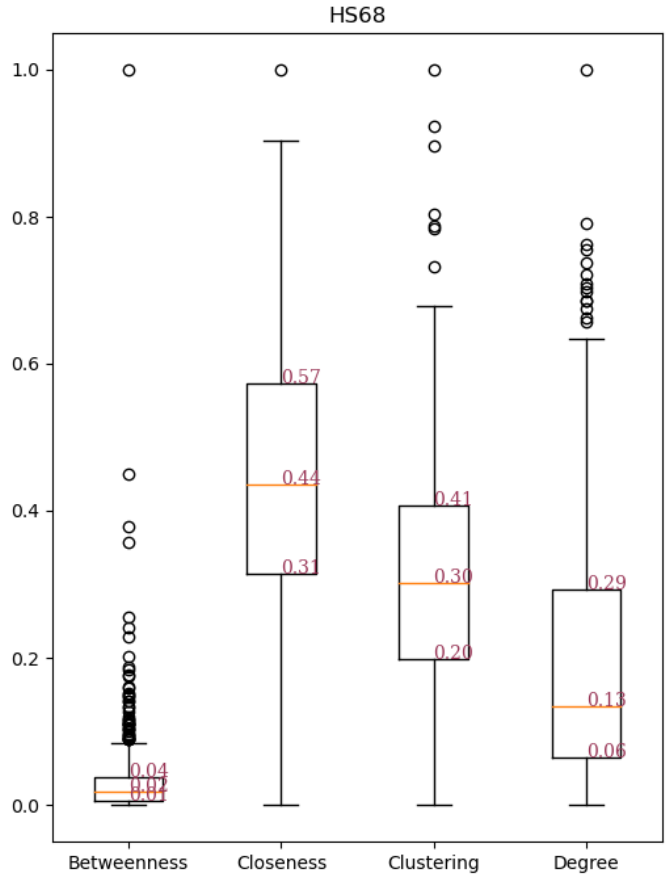

Figura 5 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um recém-nascido. Fonte: Autor.

Já para o caso da linhagem celular da pele saudável de recém-nascido, pode-se observar na figura 5 um comportamento um pouco mais consistente da medida Betweenness, embora ainda se destaque das demais medidas de centralidade. Enquanto as outras métricas se comportam de maneira semelhante às métricas da amostra A375.

- Pele saudável de paciente com melanoma no pulmão (HS895SK):

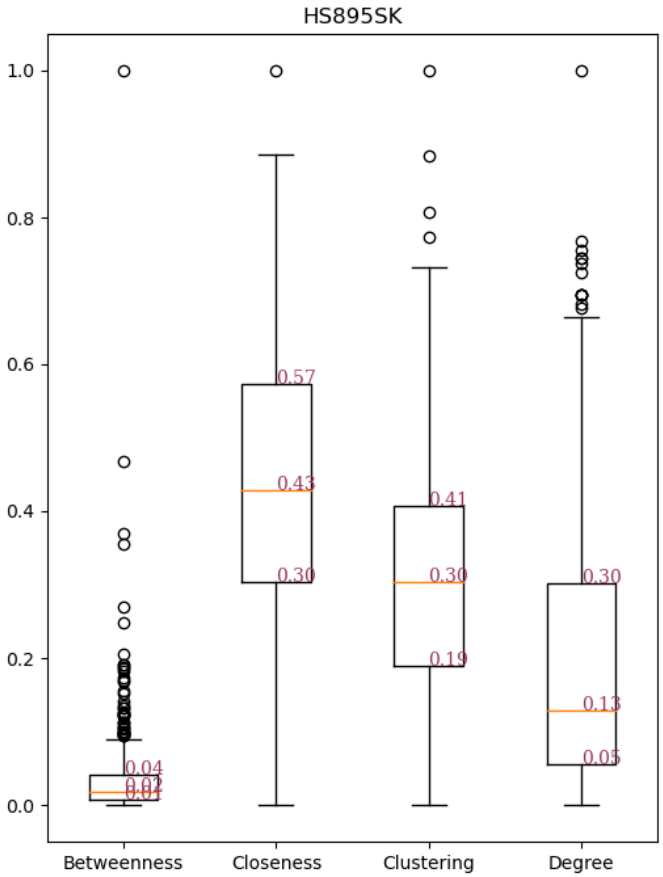

Figura 6 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um paciente com melanoma no pulmão. Fonte: Autor.

De forma geral, é possível observar na figura 6 um comportamento semelhante das medidas com as da amostra HS68, apresentando uma média de Betweennes bem baixa e destacando-se no número de outliers, enquanto a média de Closeness se aproxima de 0.43, seguida das médias de Clustering 0.30 e de Degree com 0.13.

- Local metastático do nodo pulmonar (HS895T):

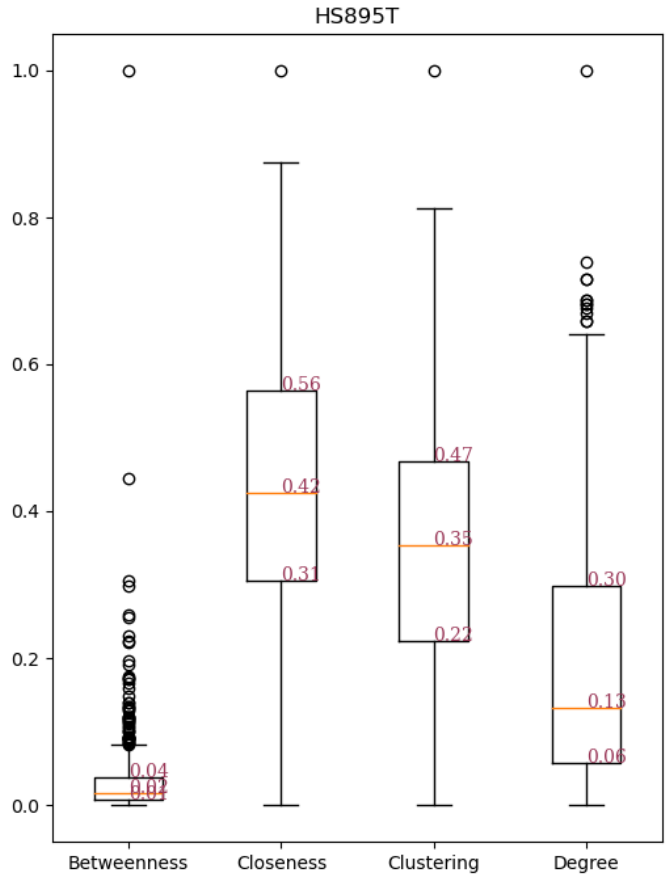

Figura 7 – Representação dos dados referentes às proteínas da linhagem celular de um local metastático do nodo pulmonar de um paciente que tinha melanoma. Fonte: Autor.

Na figura 7 é notável um comportamento bem semelhante entre as medidas de centralidade desta amostra e as outras anteriores, com exceção da média de Clustering que aumentou cerca de 0.05.

- Derrame pleural (SH4):

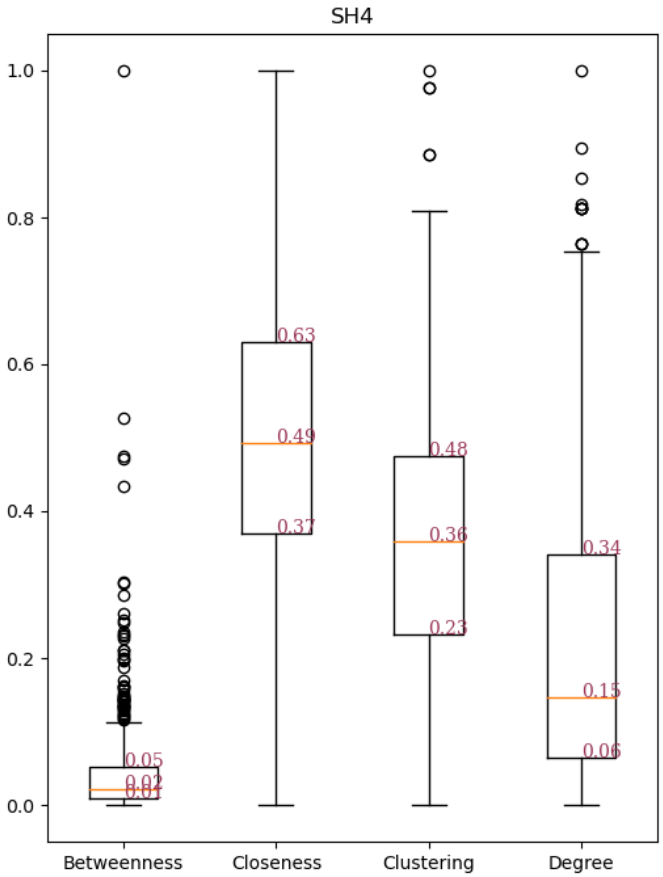

Figura 8 – Representação dos dados referentes às proteínas da linhagem celular derivada de derrame pleural de um paciente com melanoma metastático. Fonte: Autor.

Observando a figura 8, nota-se um comportamento mais semelhante entre as médias das métricas da linhagem celular de derrame pleural e as médias da linhagem celular do nodo pulmonar de um paciente que tinha melanoma, porém nenhuma grande mudança em relação ao comportamento das outras amostras, com exceção do Betweenness da amostra A375 da linhagem celular maligna do melanoma.

Analisando de modo geral, os gráficos mostraram uma grande semelhança entre as medidas das amostras, exceto pela medida de Betweenness da amostra A375 que se destacou por ter uma média bem menor em relação às outras amostras, se aproximando de zero.

Heatmap

Fonte primária do melanoma (A375):

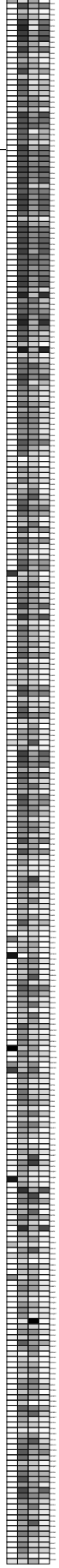

Figura 9 – Representação dos dados referentes às proteínas da linhagem celular maligna do melanoma de fonte primária. Fonte: Autor.

Pele saudável de recém-nascido (HS68):

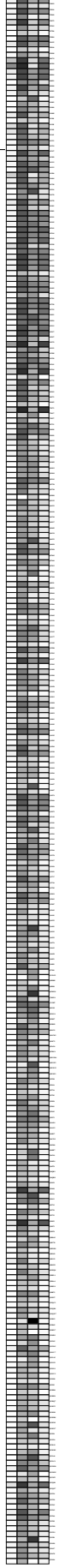

Figura 10 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um recém-nascido. Fonte: Autor.

Pele saudável de paciente com melanoma no pulmão (HS895SK):

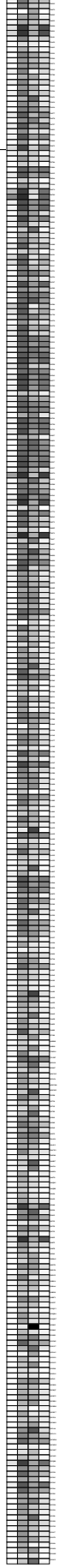

Figura 11 – Representação dos dados referentes às proteínas da linhagem celular de fibroblastos da pele saudável de um paciente com melanoma no pulmão. Fonte: Autor.

Local metastático do nodo pulmonar (HS895T):

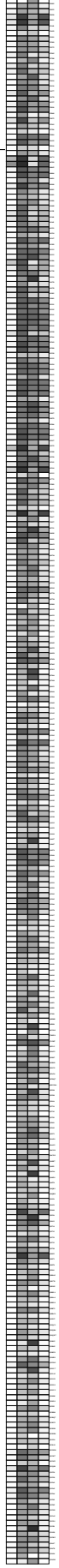

Figura 12 – Representação dos dados referentes às proteínas da linhagem celular de um local metastático do nodo pulmonar de um paciente que tinha melanoma. Fonte: Autor.

Derrame pleural (SH4):

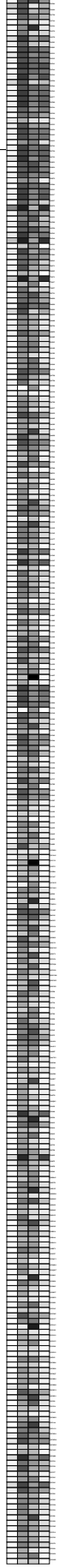

Figura 13 – Representação dos dados referentes às proteínas da linhagem celular derivada de derrame pleural de um paciente com melanoma metastático. Fonte: Autor.

Dentre as observações feitas, destacam-se que algumas linhagens têm diferenças entre a presença de algumas proteínas entre elas, porém o comportamento entre as proteínas

que coexistem tem comportamento semelhante. No entanto, houve diferença na amostra A375 em relação às outras linhagens nas proteínas P04406 (mais significativa), P07900, P34932, P50990, P50991, P60174, P60709, P62979 e P78371, onde a medida betweennes se deu bem menor e P16152 e Q5VW32 que a medida deu bem maior que as outras amostras que contém as mesmas proteínas.

Diagrama de Venn

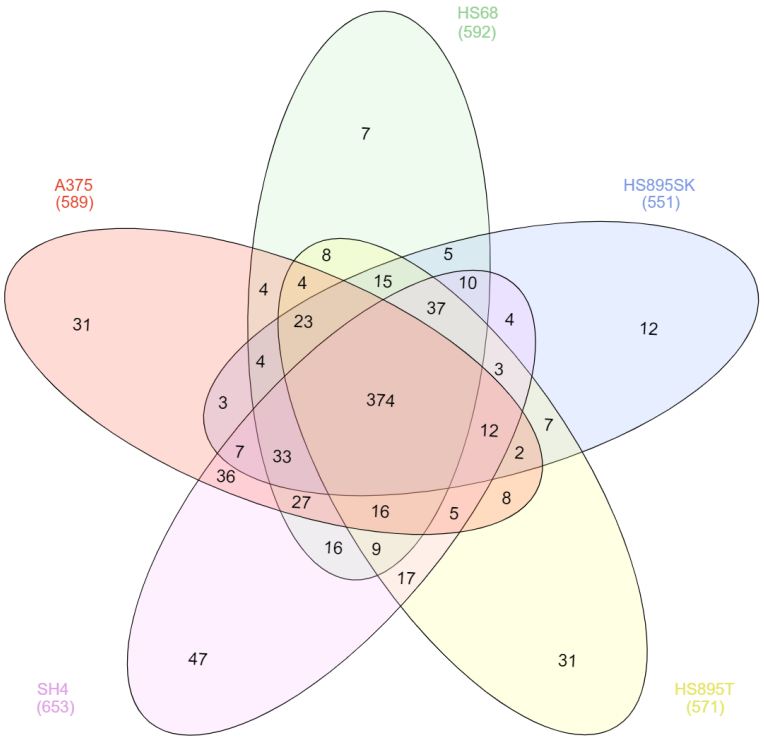

Figura 14 – Diagrama de Venn das proteínas de cada amostra dos dados. Fonte: Autor.

Analisando as quantidades de proteínas, é possível notar o número da amostra SH4, derivada do derrame pleural, possui um valor maior de proteínas diferentes em relação aos demais, além de um valor relativamente grande de proteínas semelhantes somente entre esta amostra e a da linhagem celular maligna do melanoma A375. Também é possível notar que as amostras da linhagem celular da pele saudável de um recém-nascido possui poucas proteínas exclusivas, em relação às outras amostras.

5.2 Resultado dos classificadores

K-Nearest Neighbours

Na figura 15 tem-se a matriz de confusão de cada classe classificada pelo modelo gerado pelo algoritmo KNN. Apenas pela observação, é possível notar uma facilidade um

pouco maior do modelo predizer as classes A375 e SH4 em relação às outras linhagens. Por exemplo, a classe HS68 foi mais confundida com a HS895SK do que tiveram próprios acertos. Baseando-se nesta matriz, é possível obter as medidas de avaliação geral do algoritmo para análise de qualidade do modelo. Com isto, tem-se a acurácia geral equivalendo a 0,37, precisão de 0,38 e 0,37 para Revocação e F1-Score. Pode-se inferir que a qualidade do modelo para este problema é baixa, dado que os valores das medidas de avaliação deram insatisfatórias.

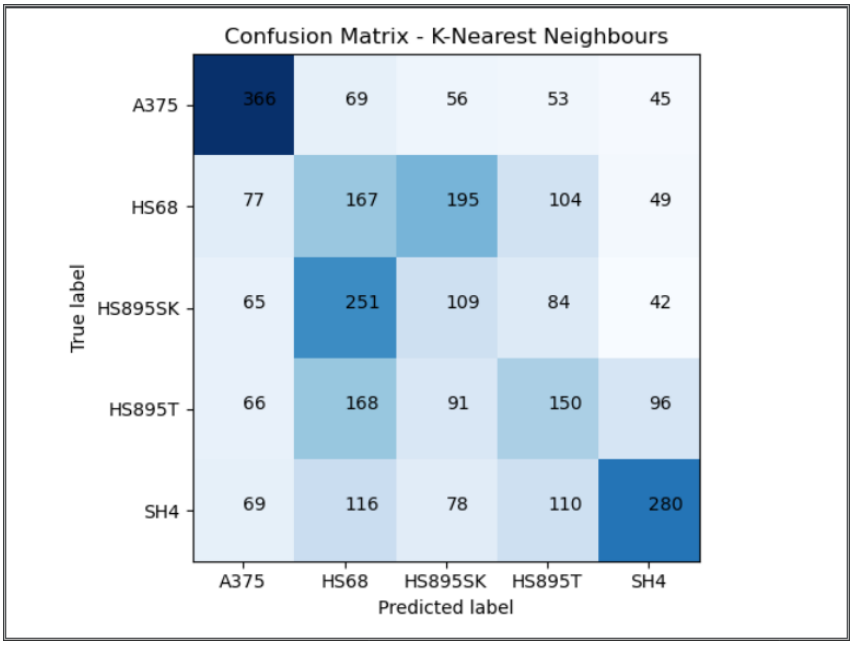

Figura 15 – Matriz de Confusão Multi-classe dos resultados do classificador K-Nearest Neighbours. Fonte: Autor.

É possível aplicar as métricas de avaliação para cada linhagem classificada pelo modelo, assim esclarecendo onde o algoritmo encontrou maior dificuldade para efetuar as predições. Assim, observa-se na figura 16 que as medidas têm grande diferença entre as classes A375 e as outras linhagens, ainda que mais próxima da classe SH4, nota-se uma dificuldade geral para encontrar um padrão na base de dados.

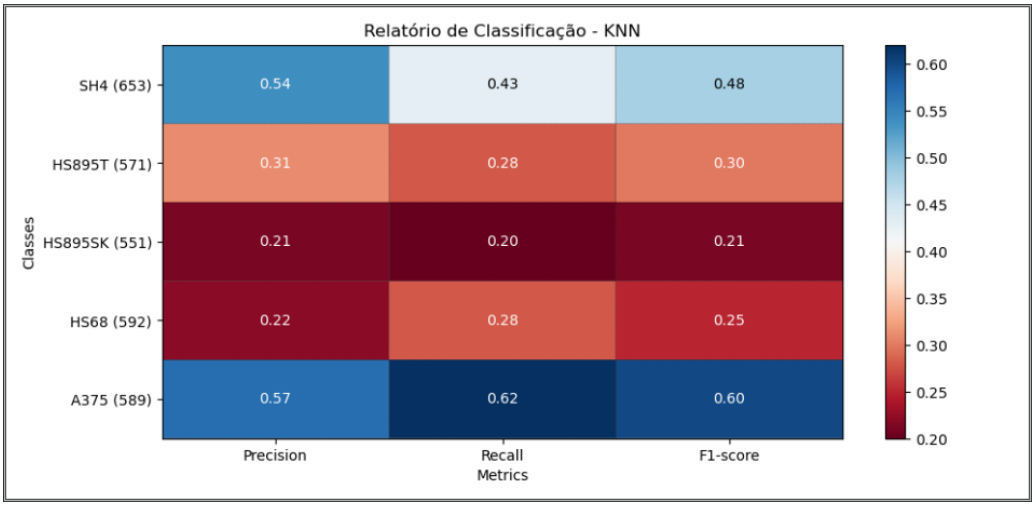

Figura 16 – Relatório de Classificação dos resultados do classificador K-Nearest Neighbours. Fonte: Autor.

Naive Bayes

Para o caso do modelo criado pelo algoritmo Naive Bayes Classifier, é possível notar na figura 17 que houve uma grande dificuldade de classificar qualquer linhagem celular da base de dados, pois a disposição dos valores preditos pelo modelo estão bem distribuídos entre todos os valores verdadeiros. A partir disto, parte-se para a análise das métricas de avaliação, obtendo uma acurácia geral de 0,24, demonstrando uma qualidade insatisfatória para o modelo inferir sobre o problema em questão. Outros valores como precisão, revocação e F1-Score também resultaram na mesma linha, sendo estes valores 0,26, 0,24 e 0,23, respectivamente.

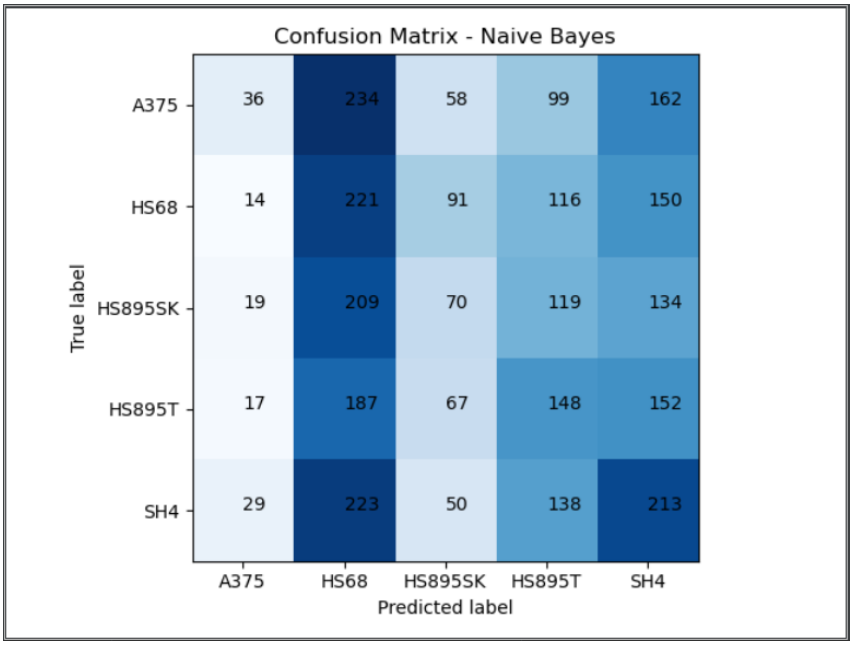

Figura 17 – Matriz de Confusão Multi-classe dos resultados do classificador Naive Bayes. Fonte: Autor.

Avaliando as classificações para cada linhagem na figura 18, encontra-se uma dificuldade grande para a classificação de qualquer classe, visto que os valores resultantes das métricas avaliativas acusam um baixo desempenho do modelo.

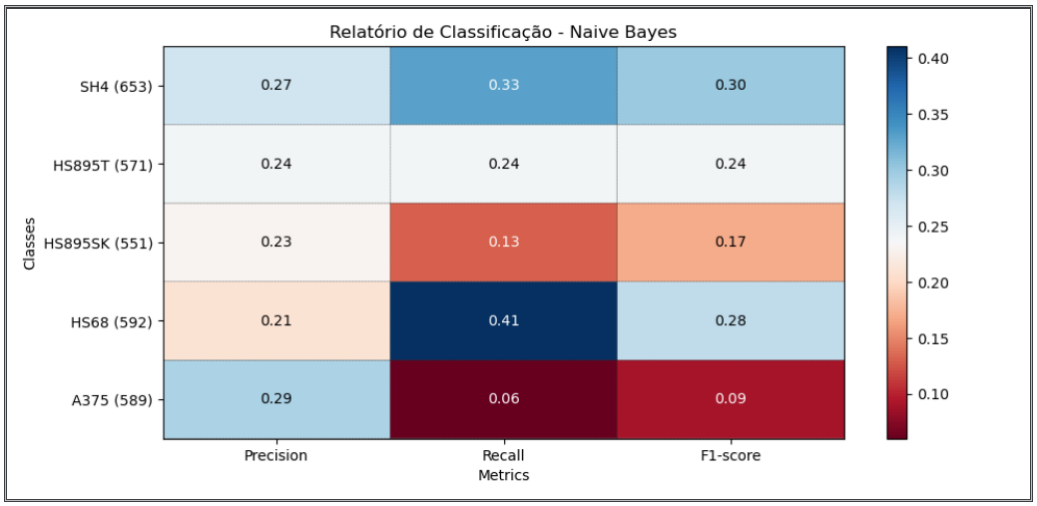

Figura 18 – Relatório de Classificação dos resultados do classificador Naive Bayes. Fonte: Autor.

Decision Tree

No caso do algoritmo Decision Tree (ou Árvore de Decisão), a figura 19 demonstra que o modelo foi capaz de classificar com facilidade a classe A375 com poucas confusões entre outras linhagens, com uma acurácia de 0,99 para esta classe em específico. De maneira mais sutil, também houve uma facilidade para classificação da linhagem SH4. Embora o

desempenho deste modelo tenha melhorado em relação aos anteriores, visto que a acurácia geral, precisão, revocação e F1-Score resultaram em 0,61, outros modelos mais robustos foram aplicados na tentativa de obter um modelo melhor para classificar a base de dados.

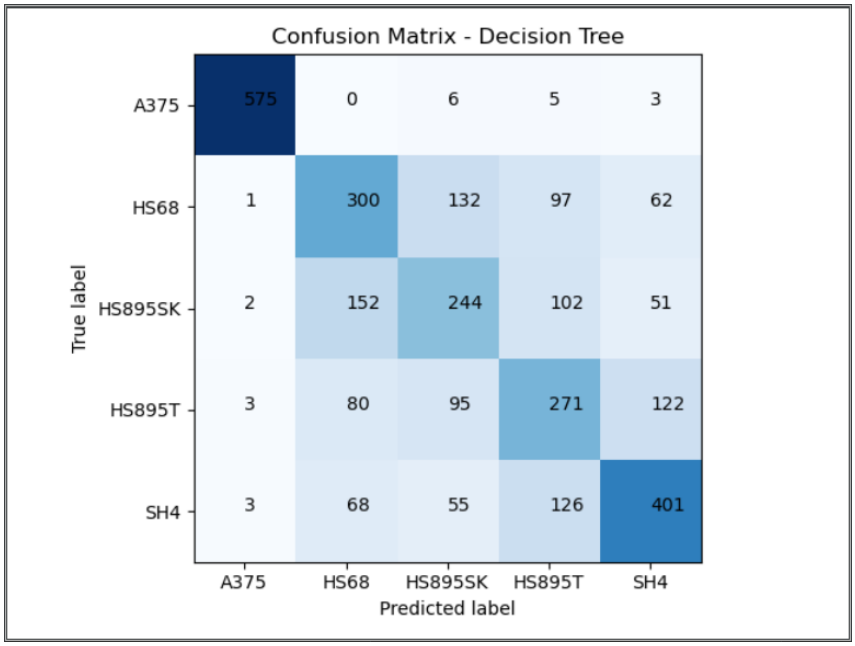

Figura 19 – Matriz de Confusão Multi-classe dos resultados do classificador Decision Tree. Fonte: Autor.

Ao aplicar a avaliação em cada classe predita pelo modelo, observa-se que, embora tenha apresentado uma precisão de 98% para predizer a linhagem A375, ainda se mantém uma dificuldade maior em classificar as linhagens HS895T, HS895SK e HS68, do que as outras classes.

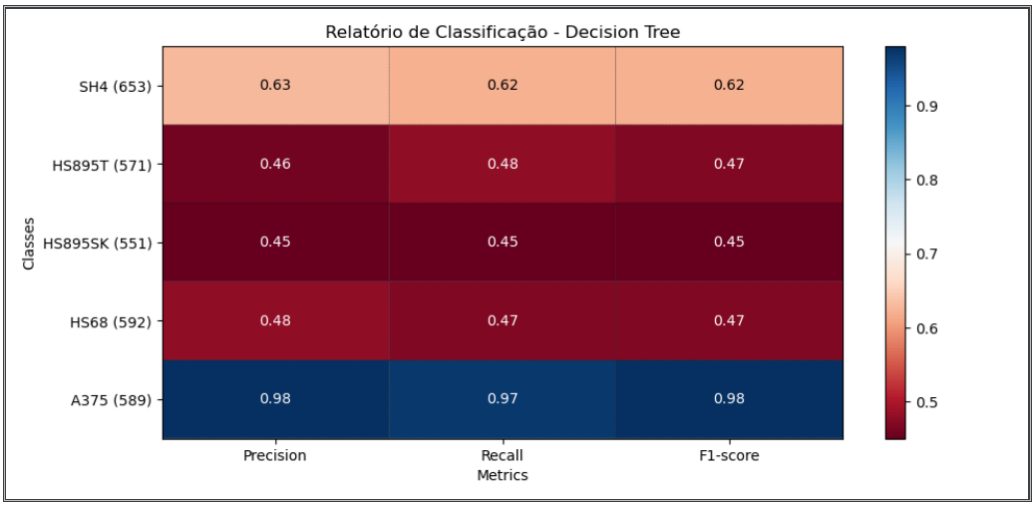

Figura 20 – Relatório de Classificação dos resultados do classificador Decision Tree. Fonte: Autor.

5.3 Resultado dos classificadores ensembles

Bagging Meta-Estimator (Decision Tree)

Analisando a figura 21, é possível notar que o desempenho de classificação do método Bagging Meta-Estimator é semelhante ao do modelo gerado pela Decision Tree, mesmo que o método Bagging esteja utilizando também o algoritmo de Árvore de Decisão. Novamente, a classificação da linhagem celular A375 se destaca pelos baixos valores de confusão, porém ao analisar as métricas de avaliação gerais do modelo, tem-se uma queda sutil de desempenho em relação ao método de Decision Tree sem a estratégia ensemble, visto que todos os resultados para a acurácia geral, precisão, revocação e F1-Score foram de 0,58, uma queda de 0,03 em relação aos resultados gerais da Decision Tree.

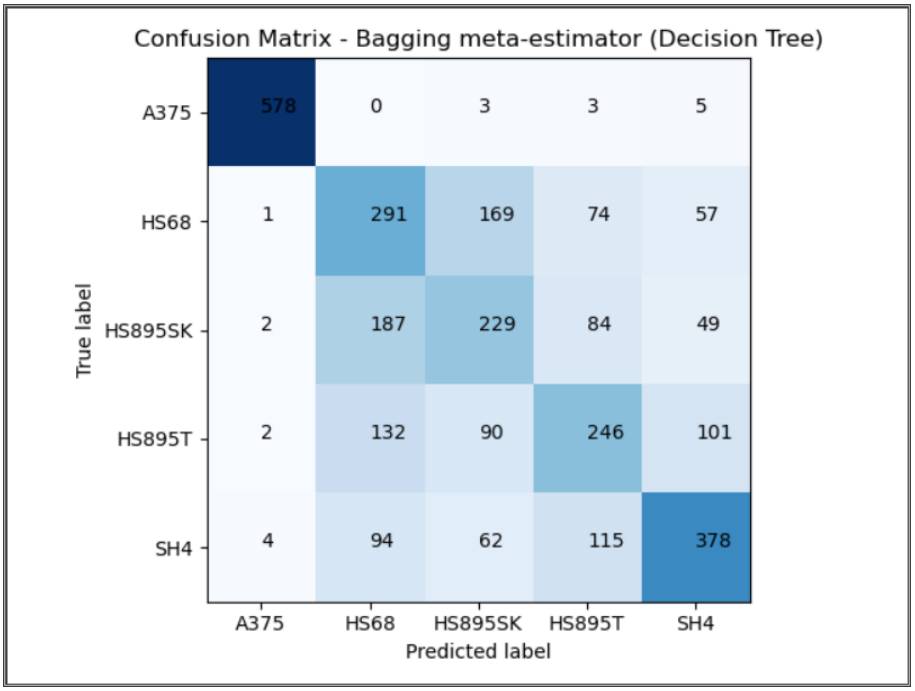

Figura 21 – Matriz de Confusão Multi-classe dos resultados do classificador Bagging Meta-Estimator (Decision Tree). Fonte: Autor.

Apesar de ser uma estratégia que visa a melhora da robustez do modelo, para este caso, pode-se observar na figura 22 que o Decision Tree com método Bagging obteve um desempenho semelhante ao Decision Tree sozinho, visto que apresenta as mesmas dificuldades e melhorias de maneira similar nos resultados avaliativos.

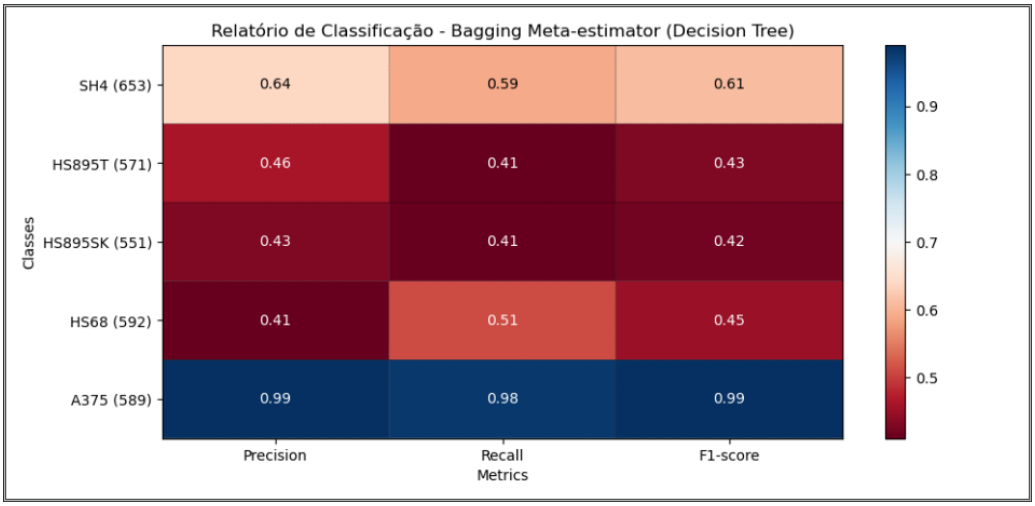

Figura 22 – Relatório de Classificação dos resultados do classificador Bagging Meta-Estimator (Decision Tree). Fonte: Autor.

Gradient Tree Boosting

Quanto à outra estratégia ensemble, no caso o método Boosting, nota-se uma grande melhora na classificação de todas as classes, com uma diminuição considerável nas confusões feitas pelo modelo. Ao analisar os valores de avaliação gerais, tem-se uma acurácia geral e uma precisão de 0,80 e uma revocação e F1-Score de 0,79, o que demonstra que o algoritmo teve uma qualidade mais satisfatória para classificar os dados e relação aos modelos anteriores.

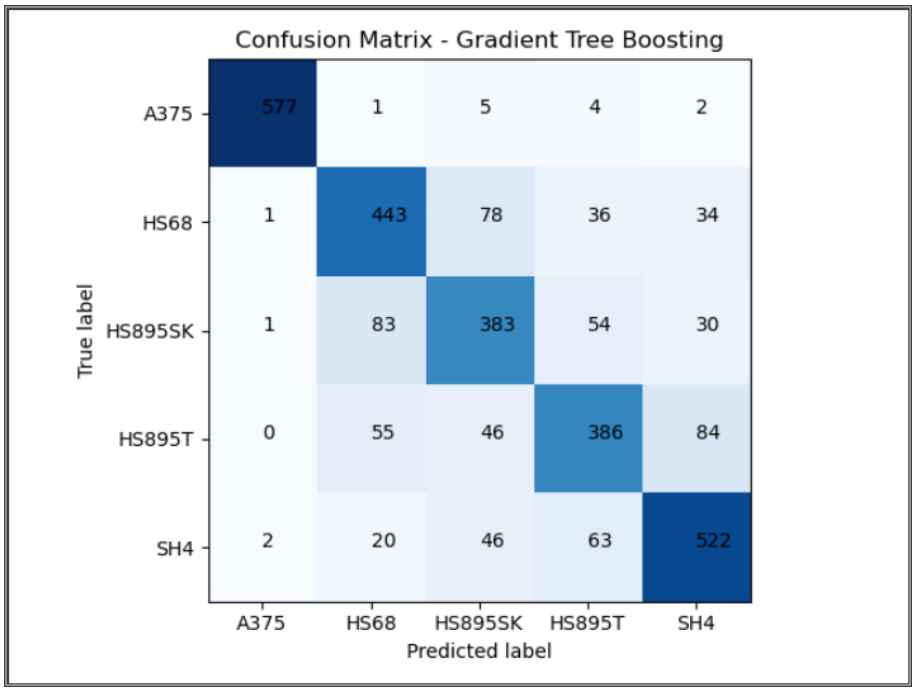

Figura 23 – Matriz de Confusão Multi-classe dos resultados do classificador Gradient Tree Boosting. Fonte: Autor.

Analisando as medidas de avaliação para cada linhagem celular classificada, nota-se uma melhora na predição de todas as classes e, apesar de ainda mostrar uma dificuldade maior para com HS895T, HS895SK e HS68, todas as precisões se mantiveram acima de 70% chegando até 100% para o caso identificar a linhagem A375. Com scores igualmente satisfatórios, o modelo se demonstrou ser de boa qualidade para o problema em questão.

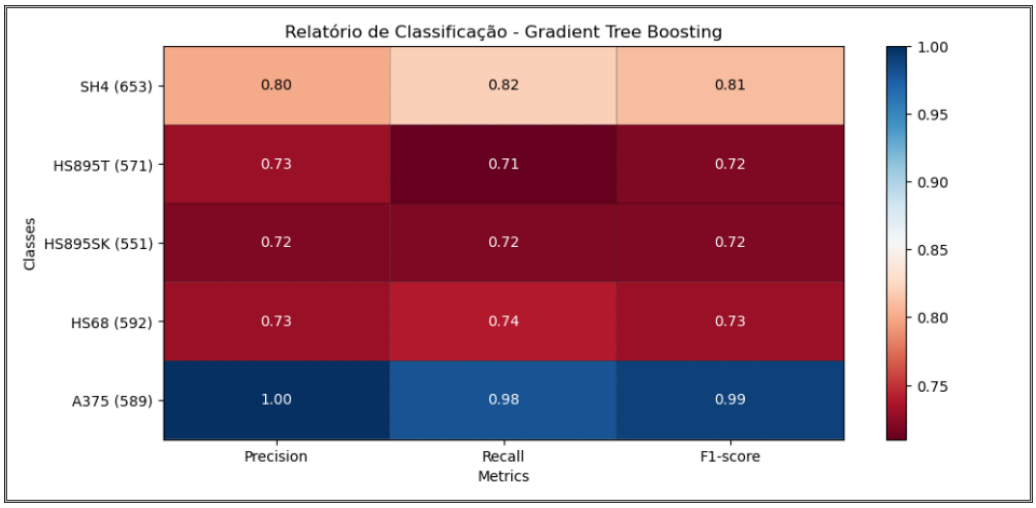

Figura 24 – Relatório de Classificação dos resultados do classificador Gradient Tree Boosting. Fonte: Autor.

Histogram-Based Gradient Boosting Classification Tree

Com uma estratégia parecida com o Gradient Tree Boosting, este algoritmo foi capaz de atingir uma classificação também semelhante quando observado na figura XX. Ainda que de maneira mais leve, a classificação teve um pequeno aumento na confusão de predições feitas. A acurácia geral, precisão, revocação e F1-Score resultaram em 0,77, demonstrando que a técnica de Boosting realmente se mostrou mais adequada para o problema em questão, quando comparada às outras técnicas aplicadas.

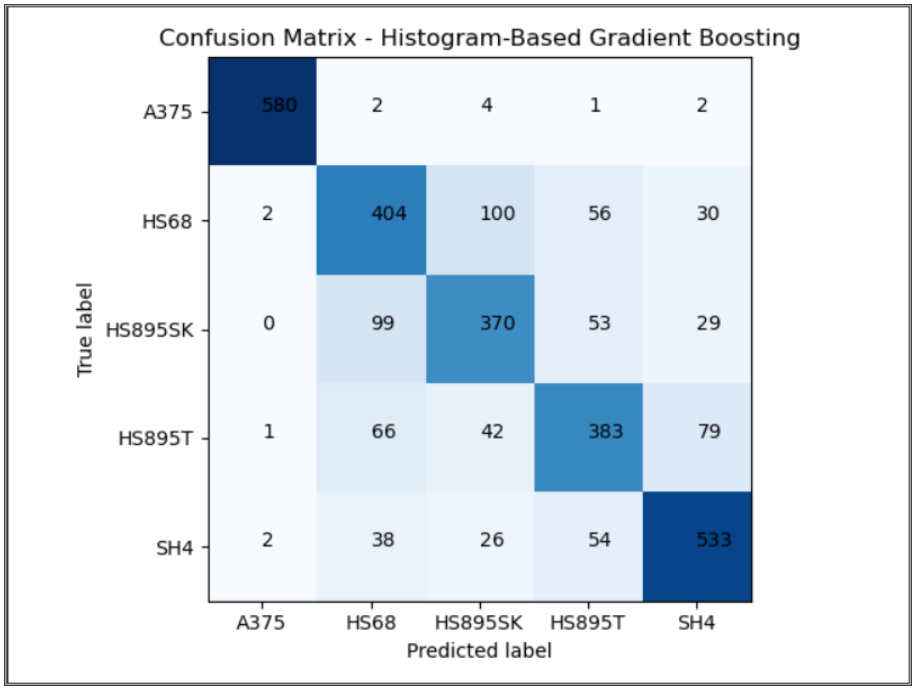

Figura 25 – Matriz de Confusão Multi-classe dos resultados do classificador Histogram-Based Gradient Boosting Classification Tree. Fonte: Autor.

Na figura 26, nota-se que a dificuldade de classificação de certas linhagens celulares se mantém, ainda que se tenham obtidos valores de precisão e score melhores. Por outro lado, o modelo chegou a um score de 0,99 para identificação da classe A375, se assemelhando aos resultados do algoritmo Gradient Tree Boosting, ainda que de forma sutil tenha encontrado uma dificuldade maior para classificação dos dados.

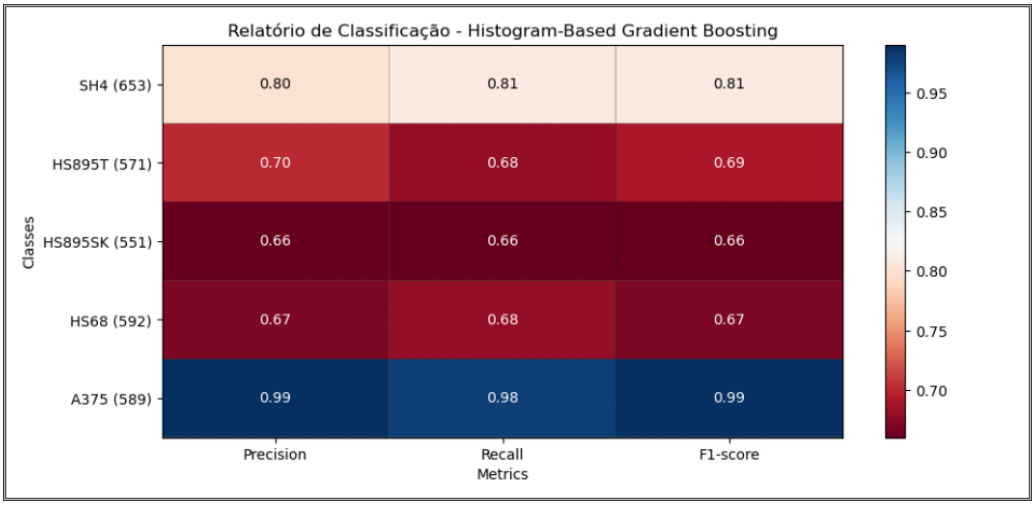

Figura 26 – Relatório de Classificação dos resultados do classificador Histogram-Based Gradient Boosting Classification Tree. Fonte: Autor.

5.4 Considerações finais

Este trabalho proporcionou ao aluno a oportunidade de se aprofundar no estudo de aprendizado de máquina, aplicado em problemas reais. Além de noções sobre o desenvolvimento de trabalhos científicos.

Os resultados obtidos mostram que os "Ensembles" obtiveram maior acurácia na classificação dos melanomas, em especial o algoritmo Gradient Tree Boosting. Desse modo, os resultados são relevantes para o estudo de melanoma e podem auxiliar no reconhecimento de melanoma em estágios iniciais e na promoção do controle da doença e qualidade de vida dos pacientes.

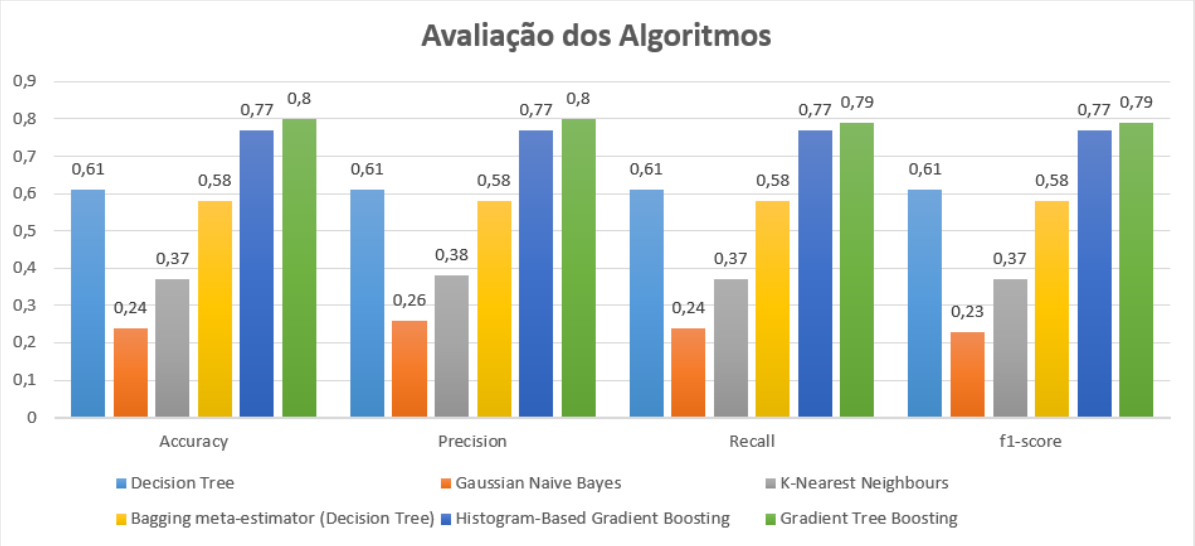

Figura 27 – Gráfico de barras apresentando as medidas de avaliação geral de todos os algoritmos em conjunto. Fonte: Autor.

Algoritmo	Acurácia	Precisão	Revocação	F1-Score
Decision Tree	0,61	0,61	0,61	0,61
Gaussian Naive Bayes	0,24	0,26	0,24	0,23
K-Nearest Neighbours	0,37	0,38	0,37	0,37
Bagging meta-estimator (Decision Tree)	0,58	0,58	0,58	0,58
Histogram-Based Gradient Boosting	0,77	0,77	0,77	0,77
Gradient Tree Boosting	0,80	0,80	0,79	0,79

Referências

1 NEWMAN, M. *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010. ISBN 0199206651, 9780199206650. Citado 3 vezes nas páginas 3, 12 e 13.

2 SHALEV-SHWARZ, S.; BEN-DAVID, S. *Understanding Machine Learning from Theory to Algorithms*. 32 Avenue of the Americas, New York, NY 10013-2473, USA: Cambridge University Press, 2014. ISBN 9781107057135. Citado 3 vezes nas páginas 3, 14 e 15.

3 FELTES, B. C. et al. *Bioinformática: da Biologia à Flexibilidade Molecular*. [S.l.]: Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 2014. ISBN 9788569288008. Citado na página 7.

4 SHOBHA, G.; RANGASWAMY, S. Chapter 8 - machine learning. In: GUDIVADA, V. N.; RAO, C. (Ed.). *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*. Elsevier, 2018, (Handbook of Statistics, v. 38). p. 197–228. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169716118300191>. Citado 2 vezes nas páginas 7 e 17.

5 CUMMINS, D. L. et al. Cutaneous malignant melanoma. *Mayo Clinic Proceedings*, v. 81, n. 4, p. 500–507, 2006. Citado na página 11.

6 SCHADENDORF, D. et al. Melanoma. *The Lancet*, v. 392, n. 10151, p. 971–984, 2018. Citado na página 11.

7 ORGANIZATION, W. H. *Radiation: Ultraviolet (UV) radiation and skin cancer*. 2017. Accessed: 2021-03-26. Disponível em: <https://www.who.int/news-room/q-a-detail/radiation-ultraviolet-(uv)-radiation-and-skin-cancer>. Citado na página 11.

8 BOCCALETTI, S. et al. Complex networks: Structure and dynamics. *Physics Reports*, p. 175–308, 2006. Citado 3 vezes nas páginas 11, 12 e 13.

9 MITCHELL, T. *Machine Learning*. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997. ISBN 0070428077. Citado 3 vezes nas páginas 14, 15 e 20.

10 HARRINGTON, P. *Machine Learning in Action*. 20 Baldwin Road, Shelter Island, NY 11964: Manning Publication Co., 2012. ISBN 9781617290183. Citado na página 14.

11 ZHU, X.; GOLDBERG, A. B. *Introduction to Semi-Supervised Learning*. [S.l.]: Morgan Claypool Publishers, 2009. ISBN 9781598295474, 9781598295481. Citado na página 14.

12 ZHOU, X.; BELKIN, M. Chapter 22 - semi-supervised learning. In: DINIZ, P. S. et al. (Ed.). *Academic Press Library in Signal Processing: Volume 1*. Elsevier, 2014, (Academic Press Library in Signal Processing, v. 1). p. 1239–1269. Disponível em: <https://www.sciencedirect.com/science/article/pii/B978012396502800022X>. Citado na página 14.

13 POLIKAR, R. et al. *Ensemble Machine Learning: Methods and Applications*. 233 Spring Street, New York, NY 10013, USA: Springer Science+Business Media, LLC, 2012. ISBN 9781441993250. Citado 2 vezes nas páginas 15 e 16.

14 DEVOS, L.; MEERT, W.; DAVIS, J. Fast gradient boosting decision trees with bit-level data structures. In: BREFELD, U. et al. (Ed.). *Machine Learning and Knowledge Discovery in Databases*. Cham: Springer International Publishing, 2020. p. 590–606. ISBN 978-3-030-46150-8. Citado na página 16.

15 HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. [S.l.]: Springer Science+Business Media, LLC, 2008. ISBN 9781282126749. Citado na página 16.

16 REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: _____. *Encyclopedia of Database Systems*. New York, NY: Springer New York, 2016. p. 1–7. ISBN 978-1-4899-7993-3. Disponível em: <https://doi.org/10.1007/978-1-4899-7993-3_565-2>. Citado na página 16.

17 TEAM, T. pandas development. *pandas-dev/pandas: Pandas*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3509134>. Citado na página 20.

18 HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 20.

19 HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 20.

20 PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 20.