UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CARLOS DANIEL ANDRADE

Estudo sobre a complexidade de dados ômicos: uma análise para a predição de sobrevida em diferentes tipos de câncer

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Profª. Drª. Mariana Recamonde Mendoza
Co-orientador: Thomas Vaitses Fontanari

Porto Alegre
2022

AGRADECIMENTOS

Agradeço primeiro Deus pela minha vida, e por todas oportunidades e desafios que me apresentou. Sou grato à minha família, em especial à minha esposa e à minha mãe, pelo apoio e compreensão. Sei que nem sempre estive tão presente quanto gostaria ou deveria. Além dos meus agradecimentos, toda a minha admiração à minha orientadora, Profª. Drª Mariana Recamonde Mendoza, por todo o apoio desde a concepção até a escrita deste trabalho. Por seus valiosos ensinamentos e incentivo. Espero poder colaborar, de alguma forma, em futuros trabalhos. Agradeço também ao meu co-orientador Thomas Vaitses Fontanari, pelas valiosas explicações, apoio na escrita e na geração das imagens. Por fim, agradeço a todas as pessoas que, de alguma forma, colaboraram comigo durante este trabalho e a minha graduação.

RESUMO

Dados ômicos fornecem uma análise em larga escala de fatores moleculares envolvidos no funcionamento dos organismos e em doenças, desde mutações até níveis de expressão para todos os genes, e têm sido amplamente explorados para o desenvolvimento de modelos preditivos de diagnóstico ou prognóstico através de aprendizado de máquina (AM). No entanto, dados ômicos apresentam uma inerente complexidade que pode dificultar o treinamento de modelos, apesar da sua riqueza de informações. Um fator crítico é o significativo desbalanceamento de classes que geralmente caracteriza dados médicos. Recentemente, outras características intrínsecas dos dados foram relacionadas ao baixo desempenho de modelos de AM para conjuntos de dados desbalanceados em geral, como sobreposição de classes (*class overlapping*) ou fronteiras de decisão complexas. Quantificar essas características através de medidas de complexidade de dados propostas na literatura permite estimar a dificuldade de um problema de classificação com base nos dados de treinamento, possibilitando um uso mais eficaz de métodos de AM. No entanto, até o momento, não houve um amplo estudo das características intrínsecas dos dados ômicos e sua correlação com a dificuldade de tarefas preditivas. Portanto, o objetivo deste trabalho é estudar as características intrínsecas de dados ômicos para análise preditiva clínica em câncer, usando medidas de complexidade de dados. Abordamos a tarefa de predição de sobrevida em 3 anos, analisando oito tipos de câncer e quatro tipos de dados ômicos (*i.e.,* variação do número de cópias (CNV), expressão gênica, expressão de microRNAs e metilação do DNA), bem como a combinação entre eles (*i.e.,* abordagem multi-ômica). Observamos que as métricas de complexidade baseadas em sobreposição de *features* e em vizinhança podem servir como preditores de desempenho, enquanto as métricas de separabilidade linear não apresentam correlação com o desempenho. Notamos uma redundância significativa entre os grupos de medidas, sugerindo possíveis simplificações e adaptações para o conjunto atualmente existente. Finalmente, nossos experimentos sugerem que os dados ômicos estudados são fortemente correlacionados em termos de complexidade dos dados, incluindo a abordagem multi-ômica. Os oito tipos de câncer demonstraram em geral bastante semelhança entre si em termos de complexidade, com o tipo ACC, no entanto, destacando-se ao apresentar dados de menor complexidade do que os outros.

Palavras-chave: Dados ômicos. métricas de complexidade de dados. aprendizado de máquina.

Study on the complexity of omics data: an analysis for cancer survival prediction

ABSTRACT

The advent of high-throughput technologies has produced a number of omics datasets that provide large-scale profiles molecular features involved in organisms' functioning and diseases, from mutations to the expression of the entire pool of genes. These resources have been widely explored to develop predictive models for clinical diagnosis or prognosis with machine learning (ML) approaches. Nonetheless, these datasets present an intrinsic complexity that may hinder model development despite their information richness. One factor is the significant class imbalance that usually characterizes omics data. More recently, however, additional data intrinsic characteristics have been linked to poor performance for learning algorithms trained with imbalanced datasets in general, such as class overlap or complex decision boundaries. Quantifying these characteristics through data complexity measures allows us to estimate the difficulty of a classification problem based on the training data, supporting a more intelligent use of ML techniques. Nevertheless, to the best of our knowledge, there has been no study on the intrinsic characteristics of omics data and their correlation with the difficulty of predictive tasks. Therefore, this work aimed to study the characteristics of different omics data commonly employed for clinical predictive analysis using a broad set of data complexity measures tailored for imbalanced domains. We focus on the task of cancer survival prediction in eight tumor types based on four types of omics data (*i.e.,* copy number variation, gene expression, microRNA expression, and DNA methylation) and the combination among them (*i.e.,* multi-omics approach). We have found that complexity metrics based on features overlap and on neighborhood can serve as good predictors of performance, while the linear separability metrics are not correlated with performance. In addition, we noticed a significant redundancy between groups of metrics, suggesting possible simplifications and adaptations for the currently existing set of data complexity measures. Finally, our experiments suggest that the studied omics data types are strongly correlated in terms of data complexity, including the multi-omics approach. In general, all eight cancer types also appeared to be highly correlated with each other, with the exception of ACC, which showed a significant lower complexity than the others.

Keywords: omics datasets, dataset complexity metrics, machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1-NN	Vizinho mais Próximo (Nearest Neighbor)
ACC	Carcinoma Adrenocortical
AM	Aprendizado de máquina
BRCA	Carcinoma Invasivo da Mama
CNV	Variação do número de cópias
COAD	Adenocarcinoma do cólon
DNA	Ácido Desoxirribonucleico
F1-MaxDR	Maximum Fisher's Discriminant Ratio
F2	Volume Overlapping Region
F3	Individual Feature Efficiency
F4	Collective Feature Efficiency
FN	Falso Negativo
FP	Falso Positivo
GLM	Generalização do Modelo Linear
K-NN	K - vizinhos mais Próximo (K-Nearest Neighbors)
KIRC	Carcinoma renal de células claras
KIRP	Carcinoma de células papilares renais
L1	Minimized Sum of Error Distance of a Linear Classifier
L2	Training Error of a Linear Classifier
L3	Nonlinearity of the Linear Classifier
LIHC	Carcinoma hepatocelular do fígado
LUAD	Adenocarcinoma do Pulmão
LUSC	Carcinoma de células escamosas do pulmão
miRNA	Micro RNA

mRNA	RNA mensageiro
MST	Minimum Spanning tree
N1	Fraction of Points on the Class Boundary
N2	Ratio of Average Intra/Inter Class NN Distance
N3	Leave-one-out Error Rate of the NN Classifier
N4	Nonlinearity of a 1-NN Classifier
NB	Naive Bayes
NN	Vizinho mais Próximo (Nearest Neighbor)
RNA	Ácido Ribonucleico
SNV	Variação simples do nucleotídeo
SVM	Support Vector Machine
T1	Fraction of Maximum Covering Spheres
TCGA	The Cancer Genome Atlas
THYM	Timoma
VN	Verdadiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

O diagnóstico e o prognóstico clínico de câncer são questões de profunda importância na medicina atual. Uma correta identificação da doença pode afetar a forma de tratamento, e até mesmo as chances de sobrevivência. Para pacientes já diagnosticados com câncer, o prognóstico de sobrevivência em si é uma questão de especial interesse. Prever a chance de sobrevivência de forma precisa pode ajudar os médicos a tomar decisões melhores quanto ao tratamento, podendo evitar excesso de tratamento e aumento de custos financeiros, conforme apontam Li et al. (2021).

É sabido que o desenvolvimento de câncer, em muitos casos, está fortemente ligado a características genéticas de um indivíduo. Esta observação torna o uso de dados do genoma e outras moléculas associadas uma importante ferramenta usada em pesquisas e em muitas outras situações, desde a prevenção até o tratamento e desenvolvimento de novas drogas. O surgimento das ciências ômicas e o avanço nas técnicas de sequenciamento e análise biomolecular gerou quantidades substanciais de informações genéticas disponíveis para atender essas novas demandas. Como consequência, houve um aumento no interesse e nas pesquisas de modelos preditivos usando tais informações.

Conforme Zhao et al. (2021), o advento da aplicação de técnicas de aprendizado de máquina (AM) na última década permitiu o desenvolvimento de classificadores para a predição de sobrevida de pacientes, treinados com conjuntos contendo atributos potencialmente preditivos. Além disso, a relativa importância de um dado preditor pode ser quantificada e, portanto sugestiva da sua relevância no prognóstico. Ambas as análises são de suma relevância no entendimento e tratamento de doenças complexas.

No entanto, conjuntos ômicos apresentam, na maioria das vezes, características intrínsecas que aumentam a complexidade dos dados para tarefas de classificação, como a alta dimensionalidade e o desbalanceamento entre as classes. A relação entre complexidade e desempenho frequentemente é objeto de estudo, e diversas pesquisas apontam para a existência de uma relação entre os dois (OKUN; PRIISALU, 2009; SOUTO et al., 2010; BOLÓN-CANEDO; MORAN-FERNANDEZ; ALONSO-BETANZOS, 2015). Tanto a mensuração da complexidade quanto a avaliação da sua correlação com o desempenho são tarefas bastante relevantes. Avaliar a complexidade de um conjunto pode ajudar a decidir se é necessário algum tratamento nos dados, ou ainda se houve redução da complexidade após o tratamento (LORENA et al., 2019). Para que se possa avaliar a complexidade de dados, Ho and Basu (2002) propuseram uma série de medidas dedicadas

a identificar diferentes características que tornam um conjunto de dados mais complexo. Posteriormente, identificou-se que tais medidas são imprecisas quando aplicadas a conjuntos desbalanceados. Assim, alguns trabalhos como o de Barella et al. (2021) apresentaram adaptações destas medidas, com ajustes para levar em conta o desbalanceamento entre classes.

O desbalanceamento dos dados ômicos em diagnóstico de câncer traz uma questão importante. Frequentemente, os problemas da área de diagnóstico e prognóstico clínico de câncer possuem interesse na classe minoritária. Imagine que um determinado tipo de câncer tenha alta taxa de sobrevida. Ainda assim, é mais importante acertar os casos em que o prognóstico de sobrevida é ruim, isto é, casos nos quais o paciente possui alta chance de óbito em decorrência da doença. Um paciente erroneamente apontado por um modelo de AM como possuindo alta chance de sobrevida após diagnóstico de câncer poderia deixar de iniciar um tratamento a tempo. Para estas situações, é importante também escolher as medidas de desempenho mais adequadas.

Um fato interessante envolvendo dados ômicos é que, para um mesmo paciente, é possível obter informações de diversos tipos de ômicas. A partir desta constatação surgem muitas possibilidades a serem exploradas. Uma delas diz respeito a buscar correlações entre diferentes tipos de ômicas, verificando se há semelhanças no comportamento dos conjuntos de dados, e se há algum deles que seja menos complexo do ponto de vista de classificação para uma tarefa de interesse que os demais. Se considerarmos que dados menos complexos estão relacionados com um melhor desempenho do modelo treinado, encontrar uma ômica para a qual é mais fácil discriminar entre classes distintas poderia ajudar na escolha do melhor conjunto de dados para solucionar uma tarefa preditiva.

Outra possibilidade é a realização de uma análise que leve em conta mais de um tipo de ômica ao mesmo tempo, isto é, uma abordagem multi-ômica. Há estudos, como o realizado por Tong et al. (2020), que apontam evidências de que a integração dos dados de múltiplos tipos de ômicas, como expressão gênica, Metilação do DNA, expressão do miRNA, e outros, podem aumentar o desempenho de modelos para predição de sobrevida. No entanto, a maioria dos estudos que abordam a questão da integração de dados multi-ômicos, avalia diretamente o desempenho, sem abordar as medidas de complexidade intrínsecas aos dados que podem nos ajudar a compreender os resultados de desempenho obtidos.

De maneira semelhante à comparação entre os tipos de dados ômicos, poderíamos realizar uma comparação entre os diferentes tipos de câncer. Mesmo que os pacientes se-

jam diferentes para cada tipo de câncer, e que sabidamente exista uma grande heterogeneidade entre tumores distintos, para um dado tipo de ômica os dados serão essencialmente os mesmos do ponto de vista conceitual e semântico. Assim, é válida a comparação para avaliar se o comportamento dos diferentes tipos de câncer é semelhante, e se há algum câncer para o qual uma determinada tarefa preditiva seja naturalmente menos complexa.

O objetivo deste trabalho é estudar as características intrínsecas dos conjuntos de dados ômicos para a tarefa de análise clinica de câncer, usando medidas de complexidade adaptadas para domínios de dados desbalanceados. Utilizamos a abordagem de prognóstico preditivo de sobrevida após três anos, analisando oito tipos de câncer e quatro tipos de dados ômicos, bem como um conjunto multi-ômico, formado pela combinação dos demais tipos ômicos. Realizamos os estudos focados principalmente na perspectiva da classe minoritária e de interesse, isto é, Sobrevida após três anos igual a "Não". Buscamos identificar medidas de complexidade que possam servir como bons preditores de desempenho neste domínio e realizamos análises comparativas entre os tipos de ômicas e entre os tipos de câncer para identificar possíveis semelhanças de comportamento, bem como conjuntos que sejam menos complexos que os demais.

Este trabalho foi organizado como segue. No Capítulo 2 apresentamos o referencial teórico, revisando conceitos de AM e de ciências ômicas relevantes para o presente trabalho. São descritas as medidas de complexidade usadas (Seção 2.1), uma visão sobre os algoritmos de aprendizado supervisionado (Seção 2.2) e uma apresentação conceitual dos dados ômicos utilizados (Seção 2.3). No capítulo 3, damos uma visão geral sobre os trabalhos relacionados. No capítulo 4, descrevemos a metodologia utilizada nos experimentos, bem como detalhamos os dados empregados no nosso estudo. No Capítulo 5, apresentamos e discutimos os resultados obtidos. No Capítulo 6, apresentamos as conclusões do trabalho e elencamos oportunidades de trabalhos futuros. Por fim, o Apêndice A apresenta figuras adicionais com resultados dos experimentos realizados.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos envolvidos no desenvolvimento deste trabalho. Iniciamos revisando as medidas de complexidade adotadas para analisar as características intrínsecas dos dados. Na sequência, apresentamos os conceitos relacionados com aprendizado de máquina supervisionado, utilizados no treinamento de classificadores para predição de sobrevida em câncer. Por fim, revisamos os tipos de dados ômicos explorados neste trabalho como atributos de modelos preditivos, contextualizando o leitor em relação ao seu formato e significado biológico.

2.1 Medidas de Complexidade de Dados

Problemas de classificação em AM podem se apresentar mais ou menos complexos, dependendo das características intrínsecas do conjunto de dados sobre o qual se trabalha. Para auxiliar na estimativa do quão difícil é a tarefa, diversas medidas de complexidade de dados foram propostas na literatura. Estas medidas de complexidades são descritores obtidos dos próprios dados usados no treinamento de modelos, que independem do algoritmo de aprendizado usado e que podem sugerir particularidades nos dados que tornam a sua classificação eventualmente mais difícil. Por exemplo, Ho (2002) aponta como fontes de dificuldade em problemas de classificação aspectos como a existência de ambiguidade entre classes, alta complexidade da fronteira de divisão que separa as classes ou existência de subgrupos nas classes, e esparsidade ou dimensionalidade dos dados. Ho and Basu (2002) compuseram um conjunto inicial de medidas práticas para quantificar as características inerentes aos dados, como a sobreposição entre valores de atributos, a separabilidade entre classes, e aspectos relacionados à geometria, topologia e densidade dos dados.

Uma contribuição interessante no estudo das características intrínsecas aos dados foi dada por Lorena et al. (2019). Neste trabalho, os autores revisaram as principais medidas de complexidade previamente propostas na literatura, estendendo o trabalho seminal de Ho and Basu (2002) ao incluir novas métricas de complexidade de dados que analisam perspectivas complementares àquelas discutidas no trabalho anterior. Um total de 22 métricas de complexidade foram discutidas e implementadas em um pacote para R, chamado ECoL (Extended Complexity Library)[1]. Adicionalmente, as métricas foram padronizadas

[1] <https://github.com/lpfgarcia/ECoL>

pelos autores de forma que as medidas fiquem no intervalo entre zero e um, e que valores mais altos representem uma complexidade maior.

Em Barella et al. (2018), os autores apontaram que muitas medidas populares de complexidade de dados possuem problemas em aferir a dificuldade de conjuntos de dados com classes desbalanceadas. Assim, os mesmos propuseram que as medidas fossem decompostas por classe para melhorar sua habilidade de medir a complexidade quando há desbalanceamento entre as classes (BARELLA et al., 2018). Estas adaptações das medidas de complexidade de dados para conjuntos com desbalanceamento entre classes foram propostas e avaliadas pelos autores, e posteriormente reunidas em um pacote R denominado ImbCoL[2] (BARELLA et al., 2021).

Para o desenvolvimento deste trabalho, foram utilizadas as medidas de complexidade disponibilizadas pelo pacote ImBCoL (BARELLA et al., 2021), e adicionalmente a medida F1 do pacote ECoL (LORENA et al., 2019), visto que a mesma não está disponível no ImBCoL. As medidas foram agrupadas em três categorias, de acordo com os trabalhos anteriores (BARELLA et al., 2021). Cada categoria busca avaliar a complexidade dos dados considerando um aspecto diferente da estrutura ou característica dos mesmos, como segue:

- Medidas de Sobreposição de *Features*: Buscam avaliar o poder de discriminação dos atributos preditivos (*i.e.,* features). Foram consideradas neste trabalho as medidas F1, F2, F3 e F4.
- Medidas de Vizinhança: Tentam caracterizar a forma das fronteiras entre as classes. Foram consideradas neste trabalho as medidas N1, N2, N3, N4 e T1.
- Medidas de Linearidade (Separabilidade Linear): Procuram avaliar o quanto as classes são linearmente separáveis. Foram consideradas neste trabalho as medidas L1, L2 e L3.

A seguir são apresentadas as medidas de complexidade originais (LORENA et al., 2019) e as respectivas adaptações realizadas por Barella et al. (2021).

2.1.1 F1 - Maximum Fisher's Discriminant Ratio

Esta medida busca calcular o quanto há de sobreposição entre a classes para cada atributo preditivo. Para isso, a medida calcula o valor de Fisher's Discriminant Ratio de

[2]<https://github.com/victorhb/ImbCoL>

cada atributo. Para um problema com duas classes, pode-se utilizar a Equação 2.1,

$$f_j = \frac{\left(\mu_{f_{c0}} - \mu_{f_{c0}}\right)^2}{\sigma^2_{f_{c0}} + \sigma^2_{f_{c1}}}. \tag{2.1}$$

Em seguida é selecionado o maior valor obtido:

$$F(T) = \max_{j=1}^{m}(f_j), \tag{2.2}$$

onde m é o numero de atributos. Com o objetivo de normalizar os valores para o intervalo $[0, 1]$, e para que valores mais altos indiquem problemas mais complexos, Lorena et al. (2019) propõem utilizar o valor de F1 com a seguinte adaptação:

$$F1(T) = \frac{1}{1 + F(T)}. \tag{2.3}$$

Uma vez que nenhuma adaptação de F1 foi apresentada por Barella et al. (2021) para dados desbalanceados, neste trabalho foi utilizada a medida como descrita na Equação 2.3.

2.1.2 F2 - Volume Overlapping Region

Esta medida calcula, para cada atributo, o volume da região de sobreposição das classes. Ou seja, procura estabelecer o tamanho do intervalo de valores do atributo que pode conter instâncias de ambas as classes, e divide pelo tamanho do intervalo total de valores possíveis.

$$F2(T) = \prod_{i=1}^{m} \frac{max(0, minmax(f_i) - maxmin(f_i))}{maxmax(f_i) - minmin(f_i)} \tag{2.4}$$

$$minmax(f_i) = min(max(f_i^{c0}), max(f_i^{c1})) \tag{2.5}$$

$$maxmin(f_i) = max(min(f_i^{c0}), min(f_i^{c1})) \tag{2.6}$$

$$maxmax(f_i) = max(max(f_i^{c0}), max(f_i^{c1})) \tag{2.7}$$

$$minmin(f_i) = min(min(f_i^{c0}), min(f_i^{c1})) \tag{2.8}$$

Um problema apontado por Barella et al. (2021) a respeito desta medida é que o valor é o mesmo para ambas as classes. Se duas classes possuírem intervalos com tamanhos

diferentes, a área sobreposta representa proporções diferentes para cada classe. Como alternativa, os autores sugerem calcular F2 para cada classe ($F2\_Partial$), utilizando apenas o tamanho do intervalo da classe de interesse no denominador.

$$F2\_Partial_{c_j}(T) = \prod_{i=1}^{m} \frac{max(0, minmax(f_i) - maxmin(f_i))}{max(f_i^{c_j}) - min(f_i^{c_j})} \tag{2.9}$$

O valor da medida será zero se ao menos um atributo não tiver regiões de sobreposição, e 1 quando as classes forem totalmente sobrepostas. Importante notar que conjuntos em que as classes apresentam fronteiras de decisão mais complexas, como quando formadas por subconceitos disjuntos, podem ser avaliados erroneamente por esta medida. Considerando como exemplo, um conjunto com o atributo x, em que a classe A apresenta valores nos intervalos $[1, 2]$ e $[3, 4]$, e a classe B apresenta valores no intervalo $]2, 3[$, F2 indicará sobreposição, mesmo que ela não exista.

2.1.3 F3 - Individual Feature Efficiency

F3 tenta avaliar individualmente a eficiência de um atributo para separar as classes e seleciona o maior valor. Lorena et al. (2019) propõem usar o complemento desta medida, calculando assim a ineficiência de um atributo para distinguir as classes. Esta abordagem faz com que valores mais altos representem uma complexidade maior. A medida divide o número de instâncias na região ambígua pelo total de instâncias, e toma o menor valor.

$$F3(T) = \min_{i=1}^{m} \frac{n_0(f_i)}{n} \tag{2.10}$$

onde:

$$n_0(f_i) = \sum_{j=1}^{n} I(x_{ji} > maxmin(f_i) \wedge x_{ji} < minmax(f_i)) \tag{2.11}$$

A medida apresenta viés para a classe majoritária. A classe minoritária poderia estar inteiramente contida na região ambígua, e ainda assim o valor calculado poderia ser baixo. A solução apresentada por Barella et al. (2021) é utilizar apenas as instâncias pertencentes à classe de interesse.

$$F3\_Partial_{c_1}(T) = min_{i=1}^{m} \frac{n_0^{c_1}(f_i)}{n_{c_1}} \tag{2.12}$$

onde:

$$n_0^{c_1}(f_i) = \sum_{j=1}^{n_{c_1}} I(x_{ji}^{n_{c_1}} > maxmin(f_i) \wedge x_{ji}^{n_{c_1}} < minmax(f_i)) \tag{2.13}$$

2.1.4 F4 - Collective Feature Efficiency

Esta medida utiliza o conceito de F3, mas tenta combinar o poder de discriminação de todos os atributos. Primeiro, encontra-se o atributo mais discriminativo usando F3, e remove-se todas as instâncias classificadas corretamente pelo atributo. Em seguida, repete-se o passo anterior até que não haja mais instâncias ou todos os atributos tenham sido usados. F4 é proporção de instâncias que não puderam ser adequadamente classificadas ao fim do processo. Formalmente, a medida pode ser calculada como:

$$F4(T) = \frac{n_0(f_{min}(T_l))}{n} \tag{2.14}$$

onde T_l é o conjunto na l-ésima iteração e $n_0(f_{min}(T_l))$ calcula o número de instâncias na região sobreposta do atributo f_{min} do conjunto T_l. Considerando a i-ésima iteração, o atributo mais discriminativo do conjunto T_i pode ser calculado com a Fórmula 2.15, onde $n_0(f_i)$ é encontrado com a Fórmula 2.11.

$$f_{min}(T_i) = \left\{ f_j \mid min_{j=1}^{m}(n_0(f_i)) \right\} T_i \tag{2.15}$$

$$T_1 = T \tag{2.16}$$

$$T_i = T_{i-1} - \{X_j \mid x_{ji} < maxmin(f_{min}(T_{i-1})) \vee x_{ji} > minmax(f_{min}(T_{i-1}))\} \tag{2.17}$$

Da mesma forma que F3, F4 possui viés para a classe majoritária. A adaptação de Barella et al. (2021) calcula a razão entre as instâncias incorretamente classificadas em cada classe e o total de instâncias da classe. $F4\_Partial$ pode ser definida com a substituição de algumas fórmulas:

$$F4\_Partial_{c_1}(T) = \frac{n_0^{c_1}(f_{min}^{c_1}(T_l))}{n_{c_1}} \tag{2.18}$$

$$f_{min}^{c_1}(T_i) = \left\{ f_j \mid min_{j=1}^{m}(n_0^{c_1}(f_i)) \right\} T_i \tag{2.19}$$

$$T_i = T_{i-1} - \{X_j \mid x_{ji} < maxmin(f_{min}^{c_1}(T_{i-1})) \vee x_{ji} > minmax(f_{min}^{c_1}(T_{i-1}))\} \tag{2.20}$$

2.1.5 N1 - Fraction of Points on the Class Boundary

N1 procura avaliar a proporção de instâncias próximas à fronteira entre classes. Para isso, a medida calcula uma *minimum spanning tree* (MST) baseada nas distâncias

entre as instâncias. Após, conta quantos vértices estão ligados a pelo menos uma instância de outra classe e divide pelo total de vértices, como definido a seguir:

$$N1(T) = \frac{1}{n}\sum_{i=1}^{n} I((X_i, X_j) \in MST \wedge y_i \neq y_j) \tag{2.21}$$

onde (X_i, X_j) representa a conexão entre as instâncias X_i e X_j, e MST é o conjunto de todas as conexões na árvore. A divisão pelo número total de instâncias do conjunto diminui o impacto da classe minoritária, gerando um viés. A adaptação de Barella et al. (2021) consiste em calcular a proporção de instâncias de uma classe que se conectam com instâncias de outra classe, gerando a medida $N1\_Partial$:

$$N1\_Partial_{c_1}(T) = \frac{1}{n_{c_1}}\sum_{i=1}^{n_{c_1}} I((X_i^{c_1}, X_j) \in MST \wedge y_i \neq c_1) \tag{2.22}$$

2.1.6 N2 - Ratio of Average Intra/Inter Class NN Distance

Esta medida compara a dispersão intraclasse e interclasses, e a razão entre elas. Para cada instância, calcula-se a distância para o vizinho mais próximo (NN) da mesma classe (intraclasse) e de classe diferente (interclasses). N2 é a razão entre a média das distâncias intra e interclasses, e apresenta valores mais altos para complexidades maiores:

$$N2(T) = \frac{\sum_{i=1}^{n} d(x_i, NN(x_i)) \in \{T \mid y = y_i\}}{\sum_{i=1}^{n} d(x_i, NN(x_i)) \in \{T \mid y \neq y_i\}} \tag{2.23}$$

Por considerar a média de todas as instâncias, esta medida também possui viés para a classe majoritária. Barella et al. (2021) sugerem corrigir este viés usando para cálculo das médias apenas os valores da classe de interesse.

$$N2\_Partial_{c_1}(T) = \frac{\sum_{i=1}^{n_{c_1}} d(x_i^{c_1}, NN(x_i^{c_1})) \in \{T \mid y = c_1\}}{\sum_{i=1}^{n_{c_1}} d(x_i^{c_1}, NN(x_i^{c_1})) \in \{T \mid y \neq c_1\}} \tag{2.24}$$

2.1.7 N3 - Leave-one-out Error Rate of the NN Classifier

Esta medida é a razão entre o número de instâncias em que o vizinho mais próximo possui classe diferente, e o total de instâncias do conjunto. É equivalente a executar o algoritmo 1-NN com o procedimento de leave-one-out, podendo ser expressa pela equa-

ção:

$$N3(T) = \frac{\sum_{i=1}^{n} I(NN(x_i) \neq y_i)}{n}, \tag{2.25}$$

onde $NN(x_i)$ é o vizinho mais próximo de x_i. Valores maiores indicam que mais instâncias estão próximas de instâncias de outra classe, sendo o conjunto, portanto, mais complexo. Se o desbalanceamento entre as classes for alto, a medida tende a se aproximar do erro médio da classe majoritária. Para corrigir este problema, Barella et al. (2021) busca avaliar uma classe por vez. Para isso, considerando apenas as instâncias da classe de interesse, divide a quantidade avaliações incorretas pelo total de instâncias da classe. Passando a apresentar a seguinte equação:

$$N3\_Partial_{c_1}(T) = \frac{\sum_{i=1}^{n_{c_1}} I(NN(x_i^{c_1}) \neq y_i)}{n_{c_1}}. \tag{2.26}$$

2.1.8 N4 - Nonlinearity of a 1-NN Classifier

Para calcular esta medida, primeiro é criado um conjunto de teste interpolando pares de instâncias da mesma classe, selecionadas aleatoriamente. A classe de cada novo ponto interpolado é a mesma das instâncias usadas na geração deste. A seguir, o conjunto original é usado como conjunto de treino em um classificador 1-NN, o qual é usado para predizer a classe das instâncias do conjunto de teste. N4 é a taxa de erro obtida, conforme a seguinte equação:

$$N4(T) = \frac{1}{l} \sum_{i=1}^{l} I(NN(x_i^{'}) \neq y_i^{'}), \tag{2.27}$$

onde l é a quantidade de instâncias interpoladas. Assim como N3, para conjuntos desbalanceados, N4 tende a se aproximar do erro da classe majoritária. A adaptação de Barella et al. (2021) consiste em considerar a taxa de erro das instâncias interpoladas da classe de interesse apenas, conforme a equação:

$$N4\_Partial_{c_1}(T) = \frac{1}{l_{c_1}} \sum_{i=1}^{l_{c_1}} I(NN(x_i^{c_1{'}}) \neq c_1) \tag{2.28}$$

2.1.9 T1 - Fraction of Maximum Covering Spheres

A medida utiliza uma abordagem topológica. O método cria hiperesferas centradas em cada instância do conjunto de dados. O raio de uma hiperesfera vai gradualmente aumentando até tocar uma hiperesfera de outra classe. As hiperesferas contidas dentro de outras maiores são eliminadas. T1 é a razão entre o número de hiperesferas restantes e o total de instâncias do conjunto. Sendo $Hiperespheres(T)$ o número de hiperesferas necessárias para cobrir o conjunto usando o procedimento acima descrito, e n o número de instâncias do conjunto, T1 pode ser expresso como:

$$T1(T) = \frac{Hiperespheres(T)}{n} \tag{2.29}$$

Se o valor de T1 for próximo de zero, significa que foram necessárias menos hiperesferas para cobrir o conjunto, que possui, portanto, menor complexidade. Porém conjuntos altamente desbalanceados e com grande sobreposição pode apresentar um valor baixo para T1, mesmo que o número de hiperesferas necessárias para descrever a classe minoritária seja próximo da quantidade de instância da mesma. Como alternativa, Barella et al. (2021) sugerem usar a razão entre a quantidade de hiperesferas necessárias para descrever a classe, e a quantidade de instâncias da classe, ficando como segue:

$$T1\_Partial_{c_1}(T) = \frac{Hiperespheres(T, c_1)}{n_{c_1}}, \tag{2.30}$$

onde $Hiperespheres(T, c_1)$ calcula o número de hiperesferas necessárias para descrever a classe c_1.

2.1.10 L1 - Minimized Sum of Error Distance of a Linear Classifier

Esta medida usa um modelo linear (como um SVM - Support Vector Machine, por exemplo) para gerar um hiperplano para separar as classes. Em seguida calcula as distâncias entre o hiperplano e as instâncias classificadas de forma equivocada. L1 é a média destas distâncias. Considerando o hiperplano gerado, L1 pode ser calculado com a equação:

$$L1(T) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mid h(X_i) \neq y_i, \tag{2.31}$$

onde ε_i é a distância entra a instância i e o hiperplano e $h(X_i)$ corresponde à classe predita pelo modelo linear. Para normalizar L1 no intervalo [0,1], e para que valores maiores representem uma complexidade maior, os trabalhos de Lorena et al. (2019) e Barella et al. (2021), e os respectivos pacotes para R, usam o seguinte ajuste para os valores de L1:

$$1 - \frac{1}{L1 + 1} \tag{2.32}$$

Como L1 apresenta viés para a classe majoritária, Barella et al. (2021) adaptam a medida para considerar apenas as distâncias das instâncias mal classificadas da classe de interesse:

$$L1\_Partial_{c_1}(T) = \frac{1}{n_{c_1}} \sum_{i=1}^{n_{c_1}} \varepsilon_i^{c_1} \mid h(X_i) \neq c_1 \tag{2.33}$$

2.1.11 L2 - Training Error of a Linear Classifier

Para calcular esta medida, primeiro um classificador linear é treinado com o conjunto de treino. O valor de L2 é a taxa de erro do classificador gerado. Valores mais altos representam bordas menos lineares. L2 é dado por:

$$L2(T) = \frac{\sum_{i=1}^{n} I(h(X_i) \neq y_i)}{n}, \tag{2.34}$$

onde $h(X_i)$ é a classe predita para a instância X_i. O trabalho de Barella et al. (2021) propõe a seguinte adaptação para a taxa de erro por classe:

$$L2\_Partial_{c_1}(T) = \frac{\sum_{i=1}^{n_{c_1}} I(h(X_i^{c_1}) \neq c_1)}{n_{c_1}} \tag{2.35}$$

2.1.12 L3 - Nonlinearity of the Linear Classifier

Esta medida usa uma técnica semelhante à medida N4. Primeiramente, cria um conjunto de teste com a interpolação de pares de instâncias da mesma classe pertencentes ao conjunto original. Em seguida, treina um classificador linear (no lugar de um 1-NN) com o conjunto original. A medida L3 pode então ser calculada pela fórmula:

$$L3(T) = \frac{1}{l} \sum_{i=1}^{l} I(h_T(x_i^{'}) \neq y_i^{'}), \tag{2.36}$$

onde $h_T(x_i^{'})$ é a predição do modelo linear treinado com o conjunto T para a instância interpolada $x_i^{'}$. A adaptação de Barella et al. (2021) para calcular a taxa de erro por classe é dada pela seguinte equação:

$$L3\_Partial_{c_1}(T) = \frac{1}{l_{c_1}} \sum_{i=1}^{l_{c_1}} I(h_T(x_i^{c_1{'}}) \neq y_i^{'}) \tag{2.37}$$

2.2 Aprendizado Supervisionado

Na área da Saúde, o desenvolvimento de modelos preditivos para diagnóstico e prognóstico clínico através do AM tem atraído o interesse de muitos pesquisadores. Dada a natureza preditiva das tarefas, a utilização de modelos de aprendizado supervisionado parece bastante adequada. No aprendizado supervisionado, o objetivo geral é utilizar um conjunto de dados rotulados, ou seja, para os quais se conhece a saída real (atributo alvo ou rótulo), para treinar um modelo que realize um mapeamento correto entre entradas e saídas e permita predizer corretamente a saída para novas instâncias, ainda não vistas (LORENA et al., 2011). Em relação ao tipo de saída a ser predita, os algoritmos de aprendizado supervisionado podem resolver dois tipos de tarefas:

- Tarefas de regressão: quando o atributo alvo é um valor numérico, podendo assumir valores em um intervalo contínuo.
- Tarefas de classificação: quando o atributo alvo é categórico, com valores pertencentes a um conjunto enumerável.

Neste trabalho, estamos interessados em tarefas de classificação. Existe uma grande quantidade de algoritmos de AM que podem ser aplicados para o treinamento de modelos de classificação, variando em relação ao tipo de representação adotada para o modelo e a estratégia utilizada para se ajustar aos dados. A seguir apresentamos os algoritmos de aprendizado supervisionado adotados neste trabalho: Naïve Bayes e Regressão Logística. Dada a ênfase deste trabalho em avaliar a complexidade na classificação derivada das características intrínseca de um conjunto de dados, optou-se por utilizar algoritmos simples em relação à existência de hiperparâmetros, isto é, algoritmos que demandam poucos ajustes na sua configuração durante o processo de treinamento.

2.2.1 Naive Bayes

O Naïve Bayes (NB) é um algoritmo de aprendizado probabilístico baseado na aplicação do Teorema de Bayes, que adota a suposição de independência condicional entre os atributos (LORENA et al., 2011). Trata-se de um algoritmo que pode ser facilmente aplicado para modelagem dos dados e de execução rápida, o que o torna um escolha interessante em muitos casos.

O Teorema de Bayes pode ser expresso pela equação:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \tag{2.38}$$

onde A e B são eventos, com $P(B) \neq 0$. $P(A)$ e $P(B)$ são chamadas de probabilidades a priori de A e B, respectivamente. $P(B \mid A)$ é a probabilidade condicional de B dado que A é verdadeiro. $P(A \mid B)$ representa a probabilidade a posteriori de observar o evento A dado que B ocorre, isto é, o grau de crença no evento A após incorporar a evidência de que B é verdadeiro.

2.2.1.1 Inferência Bayesiana

A base do algoritmo Naive Bayes é utilizar uma inferência Bayesiana para encontrar a classe a que pertence uma nova instância de entrada. Cada instância é representada por um vetor $\mathbf{x} = \{(x^j), j = 1, ..., m\}$ de m valores referentes a observações dos atributos preditivos $A_1, A_2, ..., A_m$, e a saída é representada por y. Em tarefas de classificação, os valores possíveis para y pertencem ao conjunto $\{c_1, c_2, ..., c_l\}$ de tamanho l, onde l representa o número de classes. Dada uma nova instância a ser classificada, a saída do modelo Naive Bayes é determinada encontrando-se a classe que maximiza a probabilidade a posteriori:

$$Y_{MAP} = argmax_{i=1}^{l}(P(y_i \mid \mathbf{x})) \tag{2.39}$$

Aplicando o teorema de Bayes, temos como estimar a probabilidade a posteriori de cada classe, $P(y_i \mid \mathbf{x})$, através da equação:

$$P(y_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid y_i)P(y_i)}{P(\mathbf{x})} \tag{2.40}$$

Como simplificação, é comum suprimirmos o denominador da equação considerando que o valor $P(\mathbf{x})$ será constante para todas as classes, e portanto, não afetará a

predição pelo modelo. Em relação ao componente $P(y_i)$, o mesmo pode ser suprimido quando é possível assumir que a distribuição de instâncias por classe é uniforme.

2.2.1.2 Suposição Ingênua

Ainda que se realize as simplificações mencionadas para a aplicação do Teorema de Bayes na estimativa da probabilidade a posteriori $P(y_i \mid \mathbf{x})$, o seu cálculo pode se tornar computacionalmente custoso, e até impraticável, conforme o número de atributos de **x** aumenta. Isto ocorre devido ao cálculo do componente da probabilidade condicional $P(\mathbf{x} \mid y_i)$ ser difícil de estimar visto que podem existir dependências entre os atributos de uma instância, demandando um grande número de instâncias para calculá-lo de forma precisa. Para contornar este problema, surge a ideia "Naive"(ingênua) do algoritmo, que consiste em assumir que os atributos em **x** são independentes entre si, dada a classe y (LORENA et al., 2011). Com esta suposição, tem-se que:

$$P(\mathbf{x} \mid y) = P(x^1, x^2, ..., x^m \mid y) = \prod_{j=1}^{m} P(x^j \mid y) \tag{2.41}$$

Substituindo $P(\mathbf{x} \mid y)$ na equação do Teorema de Bayes, e suprimindo $P(\mathbf{x})$, que é constante entre as classes, o cálculo de $P(y \mid \mathbf{x})$ passa a ser definido como:

$$P(y_i \mid \mathbf{x}) = P(y_i) \prod_{j=1}^{m} P(x^j \mid y_i) \tag{2.42}$$

A predição dada pelo Naive Bayes é a classe y_i que maximiza a probabilidade a posteriori conforme definido na equação acima.

2.2.1.3 Estimativa de Probabilidades e Correção de Laplace

Considerando um conjunto de dados $D = \{(\mathbf{x}_k, y_k), k = 1, ..., n\}$ com n instâncias, onde cada instância é composta por um vetor **x** de m atributos preditivos e um rótulo de saída y_k conforme definidos anteriormente, todas as probabilidades necessárias para o cálculo da Equação 2.42 a fim de classificar uma nova instância de teste t ($\mathbf{x}_t$) são estimadas a partir dos dados de treinamento. De forma mais específica:

- $P(y_i)$ é calculada como a razão entre a quantidade de instâncias que pertencem à classe y_i e o total de instâncias em D.
- $P(x^j \mid y_i)$ é calculada como a probabilidade do atributo A_j assumir o valor x^j dado

que a classe y_i é verdadeira. Há diferenças no cálculo para atributos categóricos e numéricos. Enquanto para atributos categóricos a probabilidade condicional pode ser estimada dividindo-se o número de instâncias com $x_k^j = x_t^j$ e $y_k = y_i$ pelo total de instâncias com $y_k = y_i$, para atributos numéricos é comum assumir que os valores seguem uma distribuição Gaussiana. Neste caso, é preciso estimar a média e o desvio padrão de cada classe, e a probabilidade condicional pode ser calculada como:

$$P(y_i \mid \mathbf{x}) = P(y_i) \prod_{j=1}^{m} g(x^j, \mu_{y_i}, \sigma_{y_i}), \tag{2.43}$$

onde

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.44}$$

Uma limitação do algoritmo Naive Bayes é o problema conhecido como "frequência zero". Isto ocorre quando existe um atributo A_f que, para uma determinada classe c, possua um valor (v_f) entre os seus possíveis com frequência zero no conjunto de treinamento. Como as probabilidades são calculadas a partir do conjunto de treinamento, sempre que uma nova instância apresentar o valor v_f para o atributo A_f, a probabilidade designada pelo modelo à classe c será zero. Uma solução para este problema é proposta pelo método de Correção de Laplace, que consiste em assumir a existência de uma instância a mais para cada um dos possíveis valores do atributo, em relação à classe. Embora a introdução das instâncias adicionais possa causar alguma distorção nas probabilidades calculadas, seu efeito pode ser desprezado para conjuntos de treinamento suficientemente grandes e bem como pelo fato da correção ser aplicada a todos os atributos e classes.

2.2.2 Métodos Baseados em Modelos Lineares

Um modelo linear assume que há uma relação linear entre os atributos de entrada e o atributo alvo numérico y. Ou seja, a saída pode ser obtida como a combinação linear dos atributos de entrada. Existem adaptações que podem ser feitas de modo que se possa construir relações lineares entre a entrada e a saída em alguns casos onde a relação direta não é linear, incluindo situações em que a saída é categórica.

2.2.2.1 Regressão Linear

Considere um vetor de entrada $\mathbf{x} = \{(x^j), j = 1, ..., m\}$, com m valores representando as observações para os atributos $A_1, ..., A_m$, o atributo alvo y, e um conjunto de treinamento $D = \{(\mathbf{x}_k, y_k), k = 1, ..., n\}$ com n exemplos. O modelo linear pode ser expresso pela função de saída:

$$f(\mathbf{x}) = w_0 + w_1 x^1 + ... + w_m x^m \tag{2.45}$$

onde $f(\mathbf{x})$ é a saída predita para a instância **x**, $x^1, ..., x^m$ são os atributo de entrada da instância, w_0 é o deslocamento (ou coeficiente linear da reta) e $w_1, ..., w_m$ são os pesos dos atributos $x^1, ..., x^m$ (ou coeficientes angulares da reta).

A regressão linear procura ajustar os coeficientes usados na função de predição que minimizem o erro dos valores preditos em relação aos valores reais examinando o conjunto de dados de treinamento. Para isso, usa uma função de perda (*Loss function*) que retorna uma estimativa do quão ruim foi a predição para uma dada instância. Uma *Loss function* bastante usada é o erro elevado ao quadrado. O motivo de usar o quadrado do erro, ao invés do erro simples, é evitar valores negativos:

$$Loss = (f(\mathbf{x}) - y)^2 \tag{2.46}$$

Como o objetivo é minimizar o erro do modelo como um todo, deve-se agregar os valores de cada *Loss function* individual em uma função de custo do modelo. Para isso, também é frequentemente usado o erro quadrático médio, como segue:

$$J(f) = \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 \tag{2.47}$$

Uma vez definida a função de custo, o modelo busca iterativamente ajustar os coeficientes levando a função de custo em direção ao seu mínimo.

2.2.2.2 Algoritmo do Gradiente Descendente

A função de custo acima $J(f)$ é uma função quadrática (portanto convexa), apresentando um formato de parábola e um mínimo global. Para encontrar os melhores valores dos pesos a fim de minimizar a função de custo, utiliza-se normalmente um algoritmo de otimização, sendo o Gradiente Descendente um dos algoritmos de maior sucesso para este

fim.

O Gradiente Descendente é um algoritmo que iterativamente atualiza os coeficientes da função, sendo guiado pela derivada do custo (também chamada gradiente), a qual indica a inclinação da função em determinado ponto. Assim, partindo de coeficientes iniciados aleatoriamente, o gradiente de J indica em que direção o coeficiente deve ser ajustado para levar o valor de J em direção ao mínimo local. Variando o valor do coeficiente em direção oposta ao gradiente de forma iterativa, se o passo for suficientemente pequeno, tende a convergir J para o mínimo da função. Esta análise é realizada para cada coeficiente individualmente, fixando-se os demais. Em cada passo iterativo, a atualização do coeficiente pode então ser feita subtraindo-se o valor do gradiente do coeficiente atual ajustado por uma taxa de aprendizagem α:

$$w_j = w_j - \alpha \frac{\partial}{\partial w_j} J(f) \tag{2.48}$$

onde $J(f)$ segue a equação 2.47. O objetivo da inserção de α é garantir que o passo seja suficientemente pequeno para auxiliar na convergência, evitando o fenômeno de *overshooting* (quando a atualização nos coeficientes é muito grande, fazendo com que ultrapasse o mínimo global da função). O processo iterativo é repetido até que o custo dos coeficientes seja suficiente pequeno ou algum outro critério de parada tenha sido satisfeito (por exemplo, número máximo de iterações ou redução de custo maior que um limiar pré-definido).

2.2.2.3 Regressão Logística

A regressão logística é uma adaptação que permite a utilização do mecanismo desenvolvido no modelo linear para problemas de classificação binária, podendo ser interpretada como uma instanciação da Generalização do Modelo Linear (GLM). A regressão logística consiste basicamente em adaptar a função de previsão linear do modelo de regressão linear para vinculá-la a uma função sigmoide que mapeia os valores do modelo linear para valores no intervalo [0,1]:

$$f(\mathbf{x}) = g(w_0 + w_1 x^1 + ... + w_m x^m) \tag{2.49}$$

com:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{2.50}$$

Desta forma, os valores de saída de $f(\mathbf{x})$ podem ser interpretados como a probabilidade predita para a ocorrência de um determinado evento, e podem ser utilizados para extrair uma classificação ao aplicarmos um threshold sobre os valores preditos. Por exemplo, assumindo um threshold de 0.5, para $f(\mathbf{x}) > 0,5$ a saída predita é "Sim"(1), caso contrário, a saída predita é "Não"(0).

Para um modelo de regressão logística, as funções de perda e de custo passam a ser definidas como segue:

$$Loss(f(\mathbf{x}), y_i) = y_i\,(-log(f(\mathbf{x}))) - (1 - y_i)\,(-log(1 - f(\mathbf{x}))) \tag{2.51}$$

$$J(f) = \frac{1}{n}\sum_{i=1}^{n} Loss(f(\mathbf{x}), y_i) \tag{2.52}$$

A atualização dos coeficientes da função também pode ser realizada com o algoritmo do Gradiente Descendente, aplicando-se a Equação 2.48.

2.2.3 Avaliação dos Modelos com Validação Cruzada

Quando se realiza uma tarefa de aprendizado supervisionado, um dos principais aspectos é a etapa de avaliação de modelos. A avaliação de modelos visa estimar o quão bem um modelo treinado generaliza para novos dados. Para ambas as situações - seja para gerar o modelo ou para estimar seu poder preditivo - há necessidade de um conjunto de dados. Como regra geral, os dados usados para treinamento do modelo não devem ser usados para avaliação do mesmo, sob o risco de distorcer os dados de desempenho, tornando-os muito otimistas. Assim, é importante estabelecer uma estratégia de divisão e uso dos dados disponíveis para ambos os fins.

Como o conjunto de dados disponíveis durante a criação do modelo é muitas vezes limitado, a divisão dos dados entre treinamento e teste é crítica. Se por um lado a separação de muitas instâncias para teste permite uma melhor avaliação do desempenho, por outro pode deixar poucas instâncias para o treinamento, prejudicando a o desenvolvimento de um bom modelo. Da mesma forma, reservar poucas instâncias possibilita que mais dados sejam utilizados no treinamento do modelo, mas pode limitar a capacidade de avaliar o poder de generalização do mesmo.

Uma estratégia amplamente utilizada na divisão de dados para desenvolvimento de modelos preditivos é a Validação Cruzada K-Fold (LORENA et al., 2011). A Validação

Cruzada K-Fold visa aprimorar o processo de validação de modelos ao permitir avaliar o desempenho dos mesmos com todos os dados disponíveis no conjunto de dados, bem como estimar a variação de desempenho para diferentes conjuntos de teste. A estratégia consiste em dividir o conjunto original de instâncias em K partições disjuntas e de tamanhos aproximadamente iguais. Após, seleciona-se uma destas partições como conjunto de teste, e as $K - 1$ restantes formam o conjunto de treinamento. Repete-se o processo K vezes, usando uma partição diferente como conjunto de teste a cada iteração. Ao final, estima-se o desempenho do modelo a partir da análise da distribuição de medidas de desempenho obtidas ao longo das K iterações. Assim, a Validação Cruzada K-Fold garante que todas as instâncias sejam usadas pelo menos uma vez para teste e treinamento (mas não em ambos os conjuntos ao mesmo tempo), e permite uma melhor avaliação da variância nos dados de treinamento, diminuindo a sensibilidade a um eventual viés presente no conjunto de treinamento.

É importante notar que a divisão dos dados entre os K Folds é usualmente realizada de forma estratificada. A estratificação visa permitir que os dados em cada fold, e consequentemente nos conjuntos de treino e teste, mantenham a mesma distribuição do conjunto original para o atributo alvo. Por fim, salienta-se que existe uma variação da estratégia conhecida como *Repeated Cross-Validation* (Validação cruzada repetida), que consiste basicamente em aplicar a técnica de re-amostragem aleatória na criação das partições, repetindo o processo de K-fold r vezes. Embora implique em maior custo computacional, a estratégia de Validação Cruzada Repetida permite estimar melhor o impacto na variação dos dados de treinamento sobre o desempenho do modelo.

2.2.4 Medidas de desempenho

Para problemas de classificação, existem diversas medidas que procuram avaliar o desempenho de um modelo preditivo sob diferentes aspectos, e podem se mostrar mais ou menos adequadas, a depender de características do conjunto de dados e do tipo de problema específico. Um mesmo valor de acurácia por exemplo, que reflete a proporção de instâncias classificadas corretamente pode um modelo, pode ter diferentes relevâncias para conjuntos balanceados e desbalanceados. Para problemas que possuem maior interesse na classe minoritária, esta métrica não irá refletir corretamente o desempenho.

Nesta seção, iremos apresentar as medidas de desempenho adotadas no presente trabalho. Foram selecionadas métricas capazes de melhor lidar com a classe minoritária

em problemas que apresentam desbalanceamento de classes.

Matriz de Confusão

Os resultados de teste de um modelo podem ser apresentados na forma de uma matriz, chamada de matriz de confusão. Nesta matriz, os dados são estruturados de forma a permitir entender a relação entre a classe predita e a classe verdadeira nas instâncias de teste. Importante notar que na matriz de confusão, normalmente a classe de interesse é referida como classe positiva. A Tabela 2.1 apresenta um exemplo da forma da matriz de confusão, onde:

- VP: instâncias positivas corretamente classificadas
- VN: instâncias negativas corretamente classificadas
- FP: instâncias negativas classificadas como positivas
- FN: instâncias positivas classificadas como negativas

Tabela 2.1 – Exemplo de uma matriz de confusão

		Classe Predita	
		Positivo	Negativo
Classe Real	Positivo	VP	FN
	Negativo	FP	VN

Fonte: O Autor

A partir dos dados da matriz de confusão, podemos definir um amplo conjunto de medidas de desempenho. A seguir, vamos revisar as medidas utilizadas neste trabalho: *Recall*, Precisão e F-Score.

Recall (Revocação)

Também conhecida como sensibilidade, expressa o número de instâncias da classe de interesse que o modelo prediz corretamente. Voltada para a redução de falsos negativos na classe de interesse.

$$rev(f) = \frac{VP}{VP + FN} \tag{2.53}$$

Precisão

É a proporção de acertos dentro das instâncias previstas como positivas. Voltada

para a redução de falsos positivos na classe de interesse.

$$prec(f) = \frac{VP}{VP + FP} \tag{2.54}$$

F-score

F-Score é a média harmônica entre as medidas de precisão e revocação. É uma forma de combinar ambas as medidas em apenas uma.

$$F_\beta - Score(f) = (1 + \beta)\frac{prec(f) * rev(f)}{(\beta^2 * prec(f)) + rev(f)} \tag{2.55}$$

Quando $\beta = 1$ chamamos a medida de $F_1 score$. Neste caso, a medida atribui igual importância para precisão e sensibilidade.

2.3 Dados Ômicos e o Projeto TCGA

Ciências ômicas são áreas das ciências biológicas que usam tecnologias de alto rendimento para estudar a função, diferenças e interação entre vários tipos de moléculas que constituem as células, como DNA, RNA e outras. O termo "ômica"deriva do sufixo da palavra "Genômica", tendo sido esta a primeira a surgir. Atualmente existem diversas ômicas, que surgiram com os avanços nas tecnologia de alto rendimento, como o sequenciamento, espectrometria de massa e outras. A seguir, apresentamos alguns conceitos importantes para o entendimento do presente trabalho ligados às ciências ômicas.

Genômica

A palavra foi usada pela primeira vez em 1986, pelo geneticista Dr. Thomas D.Roderick. Acredita-se que a palavra venha do termo "Genom", usado anteriormente para descrever um conjunto de cromossomos haploides. Genômica é o estudo do genoma, que inclui regiões codificantes e não codificantes do DNA. A área consolidou-se nos anos 90, com o surgimento dos primeiros projetos genoma, o que foi possível através do desenvolvimento de técnicas de sequenciamento automatizado. Uma técnica bastante usada atualmente é a tecnologia de microarranjos (microarray) de DNA. Ela permite a o sequenciamento do DNA de forma paralela, através de diversas sondas dispostas em um chip.

Segundo Cai et al. (2022), a análise genômica tem foco em variantes de nucleo-

tídeo simples, inserções e deleções, variação estrutural, e variação do número de cópias (CNV). SV são grandes variações no cromossomo, e CNV é uma forma particular de SV, e normalmente envolve amplificação ou deleção de grandes regiões do cromossomo. No presente trabalho focaremos na análise de CNVs.

Epigenômica

O epigenoma engloba modificações químicas indiretas dos nucleotídeos e proteínas que regulam a expressão gênica, sem mudar a sequência genética em si. Segundo Cai et al. (2022), a epigenômica envolve a investigação da metilação do DNA e modificação de histonas, bem como o entendimento da estrutura tridimensional do DNA. O metiloma é um dos aspectos mais bem caracterizados da epigenômica, e será o objeto de estudo no presente trabalho.

Transcriptômica

É o estudo do Transcriptoma, ou seja, o conjunto completo de todas as moléculas de RNA de uma célula. Inclui o RNA mensageiro, microRNAs (*i.e.,* pequenos RNAs não codificantes) e outros. Estudos têm usado esta abordagem para determinar como as mudanças de transcrição podem estar ligadas ao desenvolvimento de doenças. Os dados gerados são usados para analisar traços de expressão e identificar os mecanismos relacionados. Segundo alguns autores, o transcriptoma pode ser mais preditivo para respostas a drogas anti-câncer do que mutações genômicas e metilação do DNA. Conforme Cai et al. (2022), a transcriptômica analisa a abundância do conjunto completo de transcritos de mRNA, também chamado de nível de expressão gênica.

Projeto The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) é um marco na genômica do câncer. Este programa conta com dados moleculares de mais de 20 mil amostras de câncer primário e seus correspondentes normais, divididos em 33 tipos de câncer (LIÑARES-BLANCO; PAZOS; FERNANDEZ-LOZANO, 2021). A grande maioria dos dados é disponibilizado publicamente, representando um rico recursos para pesquisas biomédicas. A história do TCGA tem início em 2005, a partir de um projeto de genoma humano voltado à compreensão das alterações gênicas ligadas aos principais tipos de câncer. Desde então, o TCGA tem aumentando sua base através de diversas iniciativas, tendo gerado mais de 2,5 petabytes de dados para diversas ômicas.

O programa tem sido uma referência para a área de pesquisa, e sua base de dados tem servido de fonte para as mais diversas pesquisas, incluindo abordagens de uso de AM

para investigação de câncer (LIÑARES-BLANCO; PAZOS; FERNANDEZ-LOZANO, 2021). Parte destes trabalhos utiliza um subconjunto dos dados do projeto, realizando tratamento dos dados para seleção de atributos, remoção de instâncias com dados incompletos e outros, para então organiza-los em novas bases de dados. Duan et al. (2021), por exemplo, utilizou os dados do TCGA para produzir uma série de conjuntos de dados de quatro tipos de ômica e nove tipos de câncer com o objetivo de servirem como benchmark em futuros estudos sobre classificação de subtipos tumorais.

3 TRABALHOS RELACIONADOS

A dificuldade de problemas de classificação está muitas vezes ligada à complexidade dos dados, que surge devido a características intrínsecas destes. Diante da necessidade de critérios para avaliar aspectos que influenciam a complexidade, Ho and Basu (2002) propuseram uma série de medidas iniciais para essa tarefa, que formaram uma base para trabalhos posteriores na área. As medidas foram divididas em diferentes categorias, cada uma voltada para uma aspecto específico, como sobreposição, separabilidade linear, geometria e outros (*e.g.,* dimensionalidade). Mais recentemente, Lorena et al. (2019) realizaram uma revisão de diversas métricas de complexidade, estendendo o trabalho de Ho and Basu (2002) ao adicionar outras medidas encontradas na literatura. Além disso, Lorena et al. (2019) padronizaram as métricas colocando-as em um intervalo de $[0, 1]$, onde valores mais altos indicam maior complexidade, tornando-as assim mais facilmente comparáveis entre si e interpretáveis. No entanto, essas métricas não faziam distinção entre as diferentes classes na tarefa de classificação, levando a problemas em contextos onde as classes estavam desbalanceadas. Essa questão foi posteriormente trabalhada em uma série de artigos desenvolvidos por Barella et al. (2021), que propuseram adaptações às métricas a fim de levar em consideração o desbalanceamento entre classes existente em muitos conjuntos de dados. A seguir, revisamos os principais estudos que exploraram medidas de complexidade de dados, ajustadas ou não para o desbalanceamento de dados, em domínios similares ao abordado no presente trabalho.

Devido ao fato de o campo das ciências ômicas ser caracterizado por dados de alta dimensionalidade e significativo desbalanceamento, diversos trabalhos procuraram estudar mais a fundo outras características de complexidade que esses dados poderiam apresentar. Um dos primeiros trabalhos envolvendo análise de complexidade de dados ômicos foi desenvolvido por Okun and Priisalu (2009). Neste trabalho, foi explorada a relação entre complexidade de dados e o desempenho de modelos k-NN aplicados a conjuntos de dados de expressão gênica para a classificação binária em diagnóstico de câncer. No entanto, o trabalho não procurava caracterizar mais a fundo a complexidade dos dados e não fazia uso das medidas propostas em Lorena et al. (2019) e suas versões para dados desbalanceados Barella et al. (2021). Em Lorena et al. (2010), dados de microarranjo de expressão gênica foram usados para estudar como a escassez de dados afeta o desempenho dos classificadores em uma tarefa de seleção de marcadores gênicos. Neste contexto, escassez de dados é uma medida de complexidade dada pela razão entre a dimensionali-

dade e o número de amostras, referida como T2 em Ho and Basu (2002). As observações apontaram que técnicas de seleção de bons atributos podem reduzir os erros associados ao desbalanceamento.

Em Souto et al. (2010), os autores realizaram estudos sobre a complexidade em diagnóstico de câncer usando dados de microarranjo (*i.e.,* mensuração de expressão gênica em larga escala). Esse estudo contou com a utilização de uma parcela maior das medidas propostas por Ho and Basu (2002), e avaliou a sua correlação com os resultados da taxa de erro de um classificador do tipo Support Vector Machine (SVM). Um ponto interessante diz respeito à metodologia, onde parte dos testes foi feita sobre conjuntos onde foi introduzido ruído. A versão com ruídos foi criada a partir dos conjuntos originais com a troca dos rótulos de parte das instâncias. As análises foram realizadas para problemas de classificação binária e multiclasse e havia presença de conjuntos desbalanceados, mas estes não foram tratados.

O assunto do desbalanceamento de classes foi abordado por Lorena et al. (2012). Buscando estender seus estudos anteriores sobre a dificuldade em tarefas classificação de expressão gênica ligada a câncer, os autores apontaram que, juntamente com a escassez de dados, o desbalanceamento de classes é um tópico desafiador. A fim de estudar isso, os autores propuseram uma nova medida de desbalanceamento (C1 - razão entre número de instâncias das classes minoritária e majoritária), com o objetivo de avaliar a complexidade do conjunto também nestes termos.

Além das medidas de desempenho mais comumente usadas por diversos artigos, como acurácia e taxa de erro, há outras medidas que podem ser usadas. Algumas delas trazem informações importantes para complementar a compreensão dos dados, apresentando uma perspectiva diferente dos resultados. Por exemplo, em alguns contextos pode ser mais importante apresentar uma baixa taxa de erro relativa a uma classe específica do que uma baixa taxa de erro geral. Nesse sentido, Bolón-Canedo, Moran-Fernandez and Alonso-Betanzos (2015) procuraram estudar a complexidade de conjuntos de microarranjo de expressão gênica, levando em conta o uso de outras medidas, como sensibilidade (*i.e.,* recall) e especificidade. Um experimento interessante realizado no trabalho foi a normalização dos atributos de forma separada, após a divisão dos conjuntos de treinamento e teste.

Dando prosseguimento ao estudo de Bolón-Canedo, Moran-Fernandez and Alonso-Betanzos (2015), Morán-Fernández, Bolón-Canedo and Alonso-Betanzos (2017) realizaram uma extensa pesquisa envolvendo as medidas de complexidade e buscaram estabele-

cer alguma relação entre essas medidas de complexidade o desempenho de algoritmos de AM nas tarefas de classificação com dados de microarranjo de expressão gênica. Além disso, os autores procuraram avaliar se seleção de atributos poderia diminuir a complexidade dos dados. Mais tarde, Sánchez and García (2018) procuraram estudar as relações entre alta dimensionalidade de dados e medidas de complexidade. Para tanto, utilizaram quatro conjuntos de microarranjo de expressão gênicas.

Nenhum deses trabalhos, no entanto, buscou compreender e comparar medidas de complexidade de outros tipos de dados ômicos que não dados de expressão gênica. Além disso, os trabalhos citados foram desenvolvidos anteriormente às alterações propostas por Barella et al. (2021), que adaptavam as métricas para seu uso em dados desbalanceados. Nesse sentido, nosso trabalho se diferencia dos anteriores ao buscar estudar outros tipos de dados ômicos além dos dados de expressão gênica, incluindo também dados multi-ômicos. Não obstante, procuramos utilizar as métricas de complexidade adaptadas a dados desbalanceados, de tal forma a obter resultados mais robustos, tendo em vista o desbalanceamento frequente de dados ômicos.

4 METODOLOGIA

Neste trabalho, realizamos a análise das características intrínsecas de conjuntos de dados ômicos para análise preditiva relacionada ao prognóstico de câncer através de medidas de complexidade de dados adaptadas para conjuntos desbalanceados. Especificamente, abordamos a tarefa de predição de sobrevida em 3 anos para oito tipos de câncer, avaliando para cada um deles quatro tipos de dados ômicos e um conjunto de dados multi-ômicos (*i.e.,* representando uma concatenação dos anteriores). Para cada um dos conjuntos de dados, foram extraídas e analisadas as medidas de complexidade dos dados, bem como avaliado o desempenho na tarefa de classificação para dois tipos de modelos (*i.e.,* regressão logística e naive bayes). Os resultados obtidos foram explorados a fim de se investigar a existência de padrões de complexidade entre os dados ômicos ou tipos de câncer analisados, bem como entre características intrínsecas dos dados e a dificuldade da tarefa preditiva. Neste capítulo, detalhamos os dados e os métodos utilizados no trabalho, bem como os procedimentos realizados para geração dos resultados e as análises de correlação. Enfatizamos que todas as análises foram realizadas utilizando a linguagem de programação R, enquanto a linguagem de programação Python foi usada para algumas visualizações de dados.

4.1 Seleção e pré-processamento dos dados

Os conjuntos de dados ômicos utilizados neste trabalho derivam do artigo de Duan et al. (2021), o qual criou uma base de conjuntos de benchmark com dados ômicos relacionados a câncer para realizar uma avaliação abrangente de métodos destinados à classificação de subtipos de tumor. Quatro tipos de dados ômicos foram selecionados pelos autores e empregados neste trabalho: variação do número de cópias (CNV, de *copy number variation*) em nível de genoma, expressão de RNA mensageiro (mRNA) em nível de transcriptoma, e metilação de DNA (methy) e expressão de microRNA (miRNA) em nível de epigenoma. A escolha por estes tipos de dados se deu pelo frequente uso em estudos voltados para o diagnóstico e prognóstico de câncer a partir de dados ômicos.

Duan et al. (2021) analisaram nove tipos de câncer, selecionados por terem um número suficiente de amostras para os tipos de dados ômicos elencados acima. Dentre estes, oito foram analisados no presente trabalho: Carcinoma Adrenocortical (ACC), Carcinoma Invasivo da Mama (BRCA), Adenocarcinoma do cólon (COAD), Carcinoma de

células papilares renais (KIRP), Carcinoma renal de células claras (KIRC), Carcinoma hepatocelular do fígado (LIHC), Adenocarcinoma do Pulmão (LUAD) e Carcinoma de células escamosas do pulmão (LUSC). Os autores realizaram o pré-processamento dos dados brutos disponibilizados pelo TCGA, deixando-os preparados para análises estatísticas e computacionais. Dentre as etapas de pré-processamento aplicadas no estudo original (DUAN et al., 2021), estão a remoção de atributos ou amostras com mais de 20% dos valores faltantes, imputação de valores faltantes, remoção de *batch effects* para reduzir o efeito de aspectos experimentais técnicos no resultado das análises estatísticas, e normalização dos dados por z-scores para eliminar as diferenças devido ao uso de diferentes escalas nesses conjuntos de dados.

Por fim, dentre os conjuntos benchmark gerados pelos autores, utilizamos a versão "*Significant*", na qual mais um filtro é aplicado no pré-processamento dos dados a fim de manter somente um subconjunto de atributos significativos, isto é, aqueles que possuem maior variância entre amostras de tumor e, portanto, um potencial preditivo maior. Este processamento reduz a alta dimensionalidade inerente aos dados ômicos, ainda que os vetores de atributos resultantes para cada instância e cada tipo de dado ômico continuem com elevada dimensão. Filtrando os dados ômicos com base no desvio absoluto médio dos atributos, os dados de metilação de DNA e expressão de mRNA ficaram ambos com os top 2000 atributos, e os dados de expressão de miRNA ficaram com os top 200 atributos visto que naturalmente possuem menor dimensão que os dados citados anteriormente. Adicionalmente, para os dados de CNV, foram mantidos todos os genes em regiões genômicas com ganhos ou perdas significativas ($p-valor < 0.05$) de acordo com a ferramenta GISTIC2, podendo o número variar entre os tipos de câncer. Assim como os dados brutos gerados pelo TCGA, os dados pré-processados por Duan et al. (2021) estão publicamente disponíveis[1] e são sumarizados na Tabela 4.1.

O conjunto de dados multi-ômicos foi criado no presente trabalho para cada tipo de câncer a partir da simples concatenação entre os quatro conjuntos de dados ômicos analisados. A motivação em gerar este conjunto de dados se baseia na crescente atenção direcionada à integração de dados multi-ômicos na literatura científica. Esta abordagem tem um reconhecido potencial e valor no estudo de doenças complexas, visto que usualmente doenças como câncer são resultado de múltiplos fatores que podem causar impacto em diferentes níveis de dados biológicos, como genoma, transcriptoma e epigenoma (DUAN et al., 2021). Considerando os oito tipos de tumores e os cinco tipos de dados ômicos

[1] <https://github.com/GaoLabXDU/MultiOmicsIntegrationStudy/>

Tabela 4.1 – Número de amostras e de atributos por tipo de dado ômico utilizado neste estudo.

		Conjuntos de dados ômicos				
Câncer	**Amostras**	**mRNA**	**miRNA**	**Methy**	**CNV**	**Multi-ômicos**
ACC	77	2000	200	2000	524	4724
BRCA	759	2000	200	2000	1974	6174
COAD	291	2000	200	2000	1449	5649
KIRC	314	2000	200	2000	2102	6302
KIRP	273	2000	200	2000	1023	5223
LIHC	364	2000	200	2000	2050	6250
LUAD	450	2000	200	2000	3446	7646
LUSC	363	2000	200	2000	3074	7274

Fonte: O Autor

(incluindo o conjunto multi-ômico) explorados neste estudo, analisou-se um total de 40 conjuntos de dados diferentes durante as análises.

Para coleta de informações clínicas a respeito das amostras de tumor obtidas do TCGA e pré-processadas por Duan et al. (2021), consultamos o portal FireBrowse[2], o qual fornece acesso a uma variedade de dados genômicos e anotações sócio-demográficas e clínicas de forma simples. Em particular, obtivemos informações sobre o estado vital e o tempo de sobrevida para cada amostra de tumor contida nos dados, tendo em vista o interesse do trabalho em predizer sobrevida de pacientes com câncer. Estes dados foram registrados pelo projeto TCGA, de acordo com o seguimento feito com os pacientes com diagnóstico de câncer. Os rótulos foram sintetizados a partir destas informações clínicas a fim de gerar uma classificação binária para as amostras de câncer referente a sobrevida em 3 anos após tumor (sim ou não). O tempo de sobrevida, originalmente em dias, foi convertido para anos, e os pacientes foram classificados em dois grupos de acordo com a sobrevida em 3 anos (classe *3-years survival?*): "Yes", para pacientes que não tinham registro de óbito ou que tinham uma sobrevida superior a 3 anos, e "No", para pacientes com óbito registrado no período de 3 anos após diagnóstico de câncer.

Em cada conjunto de dados, foram mantidas apenas as instâncias que possuíam um valor válido para o rótulo, descartando-se instâncias sem esta informação. Adicionalmente, para cada tipo de câncer, todos os quatro tipos de ômicas possuíam as mesmas instâncias, garantindo uniformidade na análise das medidas de complexidade de dados. Salienta-se que a identificação única de cada amostra disponibilizada no TCGA foi usada apenas para obter uma correspondância entre amostras e análise de sobrevida (para inclusão do rótulo) e entre amostras de diferentes tipos de dados ômicos (para gerar o conjunto

[2] <http://firebrowse.org/>

Tabela 4.2 – Resumo dos dados ômicos utilizados no presente trabalho, por tipo de câncer

	Three Year Survival		
	Classe Yes	Classe No	Total
ACC	60	17	77
BRCA	707	49	756
COAD	241	48	289
KIRC	239	68	307
KIRP	246	25	271
LIHC	263	101	364
LUAD	328	116	444
LUSC	246	108	354

Fonte: O Autor

multi-ômicos), não sendo utilizada como atributo preditivo.

A Tabela 4.2 apresenta as informações básicas a respeito dos dados utilizados, especificando o número de instâncias total e por classe para cada tipo de câncer. É possível observar que com relação à quantidade de instâncias em cada valor possível do atributo alvo, os conjuntos apresentam diferentes níveis de desbalanceamento, conforme o tipo de câncer. Não obstante, a seleção dos tipos de câncer foi realizada de forma a garantir uma quantidade mínima de exemplos da classe minoritária. O objetivo principal foi evitar que fossem gerados conjuntos de dados com um número pequeno de instâncias na classe minoritária a ponto de inviabilizar a aplicação de estratégias como validação cruzada, sendo esta a justificativa para descartar um dos conjuntos de dados analisados originalmente por Duan et al. (2021) (*i.e.,* Timoma, THYM). Salientamos que embora os dados apresentem desbalanceamento entre classes, não utilizamos técnicas de pré-processamento de dados que visam mitigar este problema tendo em vista que nosso objetivo é analisar as características intrínsecas aos dados e o seu impacto na classificação.

4.2 Cálculo das medidas de complexidade dos dados

O cálculo das medidas de complexidade dos dados foi realizado na linguagem R com o auxílio dos pacotes ECoL e ImbCoL. Para cada conjunto de dados sumarizado na Tabela 4.1, foram analisadas um total de 12 medidas de complexidade dos dados, divididas em três categorias (BARELLA et al., 2021): sobreposição de *features*, vizinhança e linearidade. A relação de medidas de complexidade por categoria é detalhada a seguir.

Sobreposição de *features*:

- F1 do pacote ECoL (LORENA et al., 2019), da qual utilizamos apenas o valor da medida, sem considerar o desvio padrão fornecido pela implementação do pacote. Conforme dito anteriormente, esta medida é calculada para o conjunto como um todo, e não por classe. Passamos a nos referir a esta medida com F1-MaxDR, com o objetivo de tornar mais clara a diferenciação da medida F1-score utilizada na avaliação de desempenho dos modelos.
- F2_Partial, F3_Partial e F4_Partial do pacote ImbCoL (BARELLA et al., 2021), calculadas para cada uma das classes de saída ("Yes" e "No").

Vizinhança:

- N1_Partial, N2_Partial, N3_Partial, N4_Partial e T1_Partial, do pacote ImbCoL (BARELLA et al., 2021), calculadas para cada uma das classes de saída ("Yes" e "No").

Linearidade:

- L1_Partial, L2_Partial e L3_Partial, do pacote ImbCoL (BARELLA et al., 2021), calculadas para cada uma das classes de saída ("Yes" e "No").

Após a extração das medidas de complexidade para os dados ômicos relacionados a câncer, os resultados foram organizados em tabelas e preparados para as análises posteriores. Foram geradas tabelas com todas as medidas para ambas as classes de cada conjunto de dados, com uma medida para ambas as classes de todos os conjuntos, e com todas as medidas de uma classe para todos os conjuntos.

4.3 Treinamento e avaliação de desempenho dos modelos de classificação

Para treinamento de modelos preditivos de sobrevida em câncer, foram selecionados dois algoritmos: um classificador Naive Bayes e um classificador de regressão logística com otimização da função de custo via algoritmo do gradiente descendente. A escolha destes algoritmos se deu pela simplicidade em termos de hiperparâmetros a serem ajustados. O treinamento de modelos ocorreu para cada tipo de dado ômico, incluindo o conjunto multi-ômicos, e para cada tipo de câncer analisado. O objetivo desta análise é avaliar de forma prática a dificuldade de predizer sobrevida em 3 anos para pacientes com câncer usando cada um dos conjuntos de dados ômicos como atributos preditivos,

analisando estes resultados juntamente com as medidas de complexidade de dados.

O treinamento dos modelos foi realizado na linguagem R, utilizando o pacote *caret* (KUHN, 2008) e a função *train()*, que serve como uma interface para diversas implementações de algoritmos de aprendizado supervisionado em R. O modelo Naive Bayes foi treinado com a função *nb()* do pacote klaR, enquanto o modelo linear foi treinando utilizando-se a função *glmboost()* do pacote mboost. Os modelos foram treinados e avaliados utilizando uma Validação Cruzada K-Fold repetida. Utilizamos 5 folds e um total de 10 repetições para o processo de validação cruzada. Não foram feitos ajustes (*tuning*) adicionais nos modelos, além dos procedimentos automáticos da função *train()*. A métrica de desempenho *Recall* foi utilizada como a métrica objetivo na otimização dos modelos. Adicionalmente, mantivemos a semente do gerador aleatório adotada na partição de dados fixa, para permitir a reprodutibilidade dos resultados bem como uma comparação mais justa entre os modelos treinados.

A avaliação de desempenho de cada modelo adota como classe de interesse (comumente denominada como classe "positiva") a classe "No", a qual representa os pacientes que não sobreviveram mais de 3 anos após diagnóstico de tumor. O foco nesta classe está na importância de se identificar casos de tumor com pior prognóstico o mais breve possível, possibilitando um acompanhamento médico mais cuidadoso a fim de tentar evitar que tal desfecho se concretize. Os resultados do processo de validação cruzada k-fold repetida foram sumarizados através da análise de distribuição das métricas de desempenho ao longo dos 5 Folds e 10 repetições, sendo analisadas, portanto, um total de 50 amostras de desempenho. Os resultados também foram visualizados na forma de uma matriz de confusão agregada, a qual mostra as contagens agregadas de amostras, isto é, para todas as reamostragens realizadas, em cada uma das células da matriz de confusão. Dessa forma, as medidas obtidas a partir da matriz de confusão representam o desempenho médio para o processo de validação cruzada k-fold repetida. As medidas de desempenho obtidas foram reunidas em duas tabelas para análise posterior dos resultados: uma tabela com os resultados para cada modelo, e outra tabela com os valores para todos conjuntos de dados analisados (*i.e.,* tipos de dados ômicos e tipos de câncer).

4.4 Análises de Correlação

Nos problemas envolvendo conjuntos desbalanceados, é bastante comum que o interesse seja maior na classe menos frequente, em especial nas aplicações médicas voltadas

para diagnóstico clínico. Tendo isso em mente, as análises foram, em grande parte, voltadas para a classe minoritária. Também buscou-se identificar propriedades que pudessem apontar um caminho para melhorar a qualidade das predições. Para este fim, investigamos a correlação de Pearson entre as medidas realizadas, usando matrizes de correlação, e algumas outras características relevantes. No próximo capítulo iremos detalhar os resultados das análises de correlação, e outros achados.

5 RESULTADOS

Neste capítulo, apresentamos os resultados obtidos durante o desenvolvimento deste estudo. Primeiramente, serão apresentadas as medidas de complexidades de dados para os conjuntos de dados ômicos analisados neste estudo, bem como o desempenho dos modelos de classificação na predição de sobrevida em 3 anos. Após a discussão destes resultados básicos, passamos para a interpretação dos dados a fim de investigar a existência de padrões nas medidas de complexidade de dados ômicos ou de associações entre medidas de complexidade e desempenho de modelos preditivos. Conforme explicado anteriormente, a maior parte das análises foi feita a partir da classe minoritária ("No"), que é a classe de interesse na predição de sobrevida em câncer. Portanto, convencionamos que neste capítulo, exceto quando explicitamente informado o contrário, os dados e as medidas dizem respeito à classe de interesse apenas.

As seções a seguir apresentam os resultados de nossas análises visando responder às seguintes perguntas:

- Como as medidas de complexidade de dados variam para os diferentes tipos de dados ômicos e de câncer analisados? (Seção 5.1)
- Como é o desempenho de modelos Naive Bayes e Regressão Logística na tarefa de predição de sobrevida em 3 anos para os tipos de câncer analisados? (Seção 5.2)
- Como as medidas de complexidade se correlacionam entre si e com as métricas de desempenho? (Seção 5.3)
- Entre os tipos de dados ômicos analisados, há semelhança quanto aos padrões de comportamentos em termos de complexidade e desempenho preditivo? Há algum tipo de dado ômico que seja menos complexo para um ou mais tipos de câncer? (Seção 5.4)
- Entre os tipos de câncer, há grupos com comportamentos semelhantes? Há algum câncer para o qual a predição de sobrevida seja mais fácil? (Seção 5.5)

5.1 Apresentação das Medidas de Complexidade dos Dados Ômicos

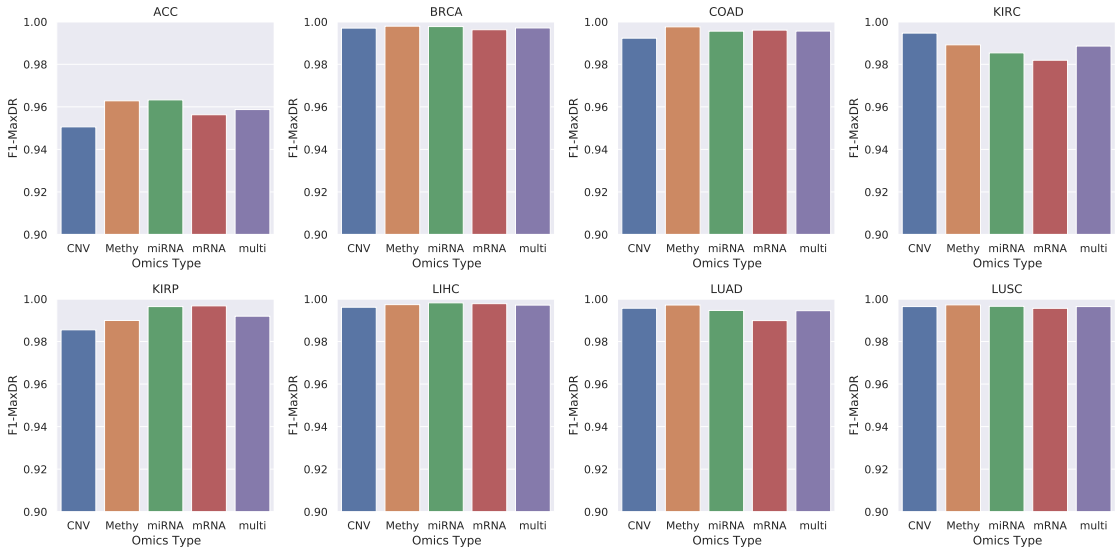
Conforme descrito no Capítulo 4, calculamos as medidas de complexidade citadas para cada um dos conjuntos de dados ômicos nos oito tipos de câncer abordados neste trabalho. A seguir, apresentamos observações gerais sobre as medidas de complexidade

e algumas comparações entre elas. Devido ao grande volume de informações geradas, nos limitamos a incluir nesta seção apenas os gráficos relacionados a algumas medidas de complexidade que sumarizam os padrões observados. A lista completa dos gráficos para todas as medidas de complexidade calculadas pode ser consultada no Apêndice A.

As Figuras 5.1, 5.2, 5.3, 5.4, 5.6 e 5.7 mostram os gráficos para a complexidade de dados em termos de F1-MaxDR, F2_Partial, F3_Partial, N1_Partial, N3_Partial e L1_Partial, respectivamente. Cada gráfico contém os valores da medida para um dos tipos de câncer, e cada barra se refere a um tipo de ômica.

Em relação às medidas de sobreposição de *features*, conforme se observa nos gráficos para F1-MaxDR (Figura 5.1) e F3_Partial (Figura 5.3), dado um tipo de câncer, o nível de complexidade é alto e bastante próximo para todos os tipos de ômicas. Em ambas as medidas, o câncer ACC foi o que obteve valores menores, indicando que a tarefa de predição de sobrevida em 3 anos a partir de dados ômicos para este tipo de tumor é um pouco menos complexa. Observamos também algumas variações em relação aos tumores KIRC e KIRP: enquanto no primeiro os dados de mRNA apresentaram os menores valores de F1-MaxDR, no segundo a menor complexidade de acordo com a mesma métrica foi encontrada com os dados de CNV. Para F3_Partial, exceto para KIRC e LUSC, os demais tipos de câncer apresentaram valores mais altos para mRNA, embora no geral os valores sejam bem próximos entre as ômicas.

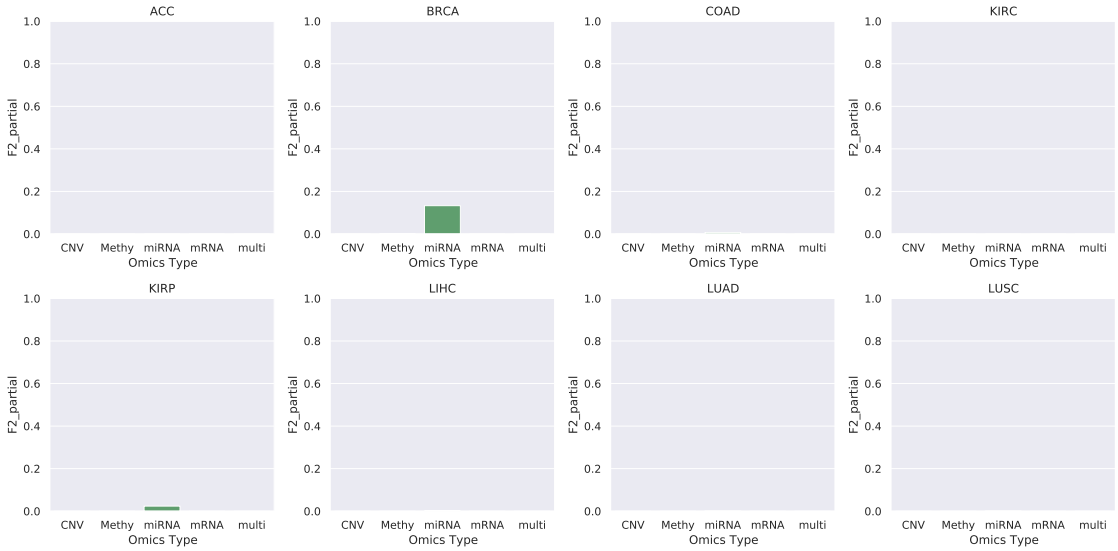
Figura 5.1 – Análise da medida de complexidade F1-MaxDR, por tipo de câncer e tipo de ômica



Fonte: O Autor

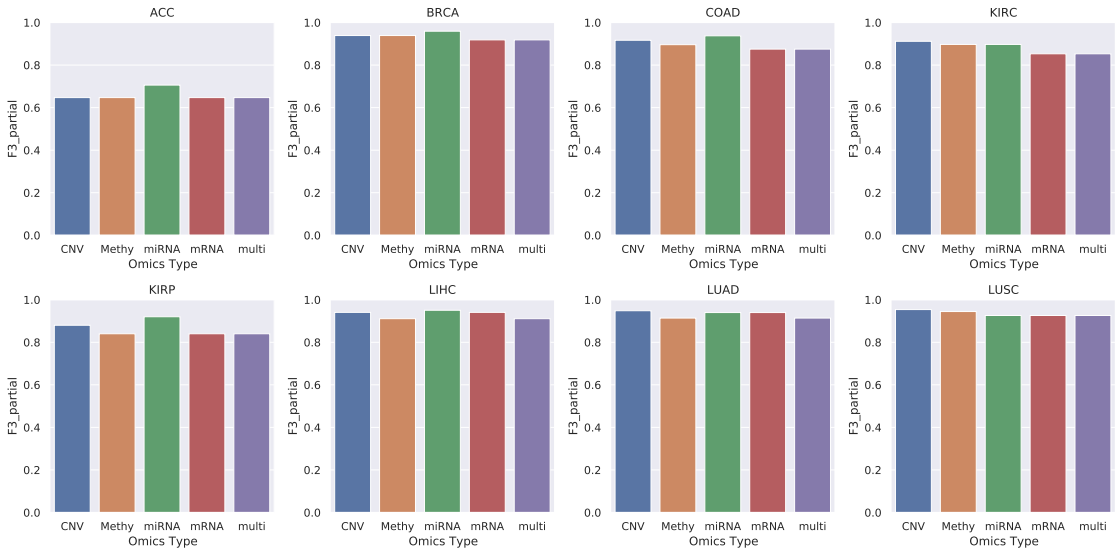
A medida F2_Partial (Figura 5.2) apresentou valores extremamente baixos e com

Figura 5.2 – Análise da medida de complexidade F2_Partial para a classe "No", por tipo de câncer e tipo de ômica



Fonte: O Autor

Figura 5.3 – Análise da medida de complexidade F3_Partial para a classe "No", por tipo de câncer e tipo de ômica



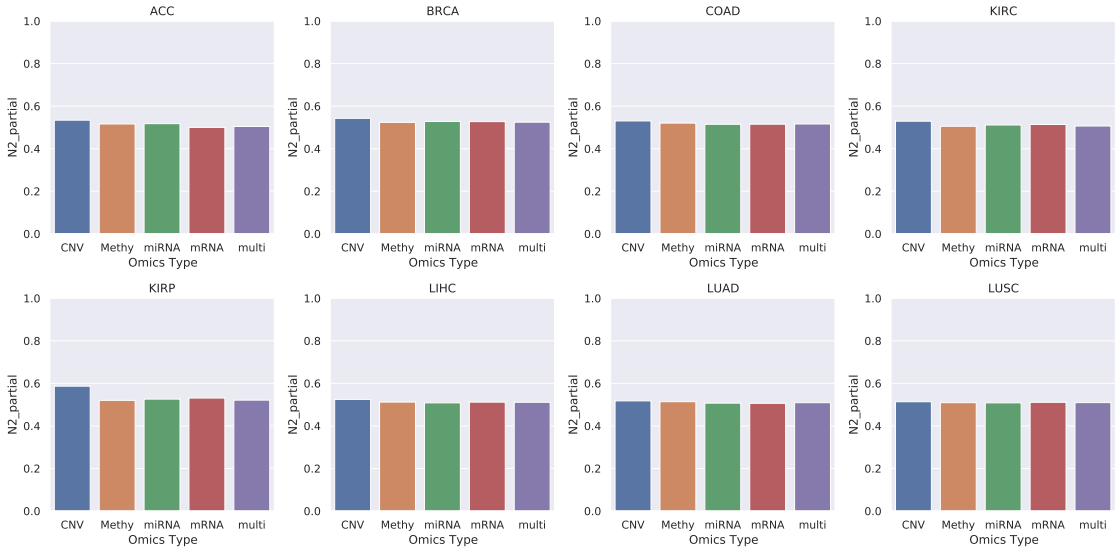
Fonte: O Autor

grande variação de escala, com a magnitude dos resultados variando de 10E-1 a 10E-214, tornando impraticável uma avaliação direta usando esta medida. Por esta razão, não utilizamos a medida para maiores observações, apesar de termos mantido nas análises posteriores de correlação. Um provável motivo para esta amplitude na escala de valores está ligado à alta dimensionalidade dos conjuntos avaliados. É possível que um grande

número de atributos com área de sobreposição menor que 1 tenha potencializado a redução de escala. Ajustes no cálculo da medida para compensar estas distorções poderiam ser objeto de estudos futuros.

Outra observação interessante é que F3_Partial e F4_Partial (Figura A.6), apresentaram não só alta correlação entre si, como discutiremos mais adiante, mas também valores idênticos em todos os conjuntos de dados analisados. Este comportamento surgiu provavelmente porque F4_Partial aplica iterações sucessivas de F3_Partial. Isso sugere que para conjuntos com características semelhantes às dos dados estudados, F3_Partial e F4_Partial são redundantes. Considerando que o custo computacional de F4_Partial é maior, devido à repetição de F3_Partial, seria razoável não realizar o seu cálculo nessas situações. Outro argumento para suprimir F4_Partial, é que a presença de medidas redundantes pode influenciar na avaliação das observações, pois de certa forma, dobra o peso da característica mensurada.

Figura 5.4 – Análise da medida de complexidade N2_Partial para a classe "No", por tipo de câncer e tipo de ômica

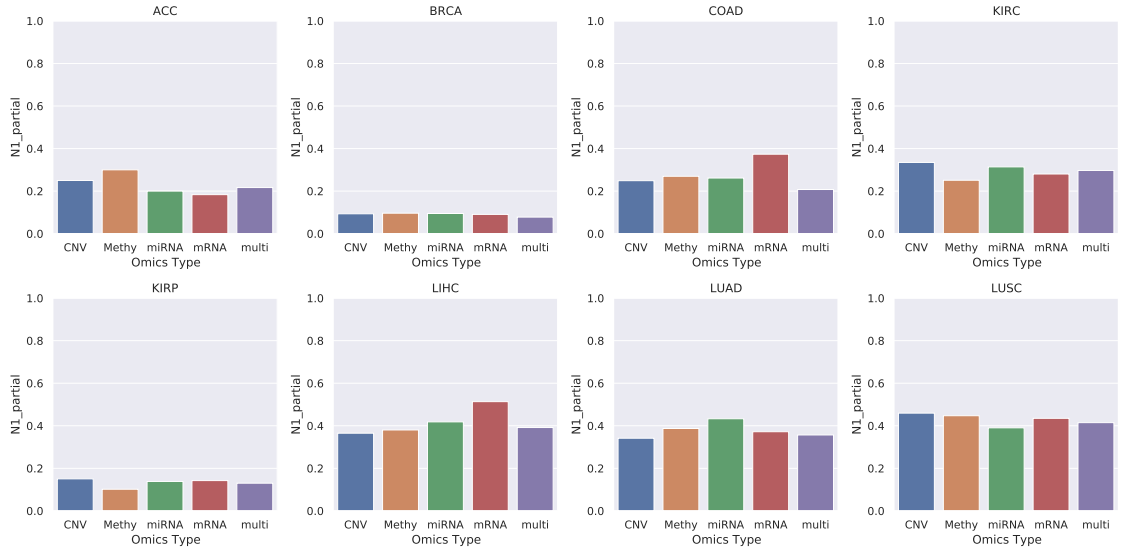


Fonte: O Autor

A medida de vizinhança N2_Partial (Figura 5.4) apresenta valores muito semelhantes para os diferentes tipos de ômicas em todos os tipos de tumores. Em alguns casos, como ACC, KIRC e KIRP, as diferenças acentuam-se um pouco para dados de CNV, que apresentam um aumento nos valores de N2_Partial. Também observamos valores mais altos da medida N4_Partial (Figura A.14) para o CNV em relação aos demais tipos de dados ômicos, sendo as diferenças muito mais significativas para esta medida de complexidade. Esta tendência de maior complexidade dos dados de CNV de acordo com N4_Partial foi

observada em todos os tipos de tumores. Uma grande semelhança entre ômicas e tumores também foi observada para a métrica T1_Partial (Figura A.16), na qual todos os conjuntos de dados possuem valores muito próximos a 1.0, com exceção dos dados de CNV para ACC. Ou seja, este caso em particular demanda um menor número de hiperesferas para cobrir todo o conjunto de dados da classe "No" (o mesmo achado foi encontrado para a classe "Yes", como mostrado na Figure A.17).

Figura 5.5 – Análise da medida de complexidade N1_Partial para a classe "No", por tipo de Câncer e tipo de Ômica

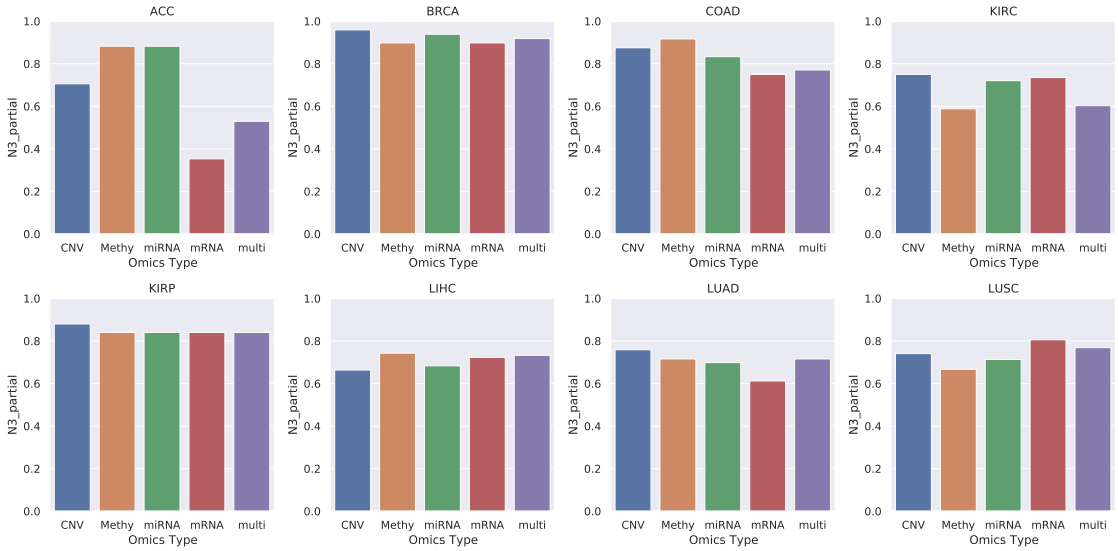


Fonte: O Autor

As medidas de vizinhança N1_Partial (Figura 5.5) e N3_Partial (Figura 5.6) para a classe "No" apresentam padrões mais variados entre os diferentes tipos de câncer e de dados ômicos. Podemos observar que de acordo com N1_Partial, que assume valores mais altos para fronteiras de decisão mais complexas, a complexidade é maior para LIHC, LUAD e LUSC, e menor para BRCA e KIRP. Entretanto, pelo critério N3_Partial, observamos que as instâncias da classe "No" em BRCA e KIRP tendem a ter valores mais altos, indicando que estão próximas de exemplos da classe "Yes", dificultando sua predição.

Por fim, discutimos as medidas de Linearidade L1_Partial, L2_Partial e L3_Partial, as quais quantificam o quanto as classes são linearmente separáveis. A três medidas de linearidade apresentaram comportamentos muito semelhantes entre si, portanto, apresentamos apenas a L1_Partial na Figura 5.7 (os resultados para L2_Partial e L3_Partial podem ser consultados no Apêndice A). Conforme pode ser visto na Figura 5.7, a complexidade calculada foi zero para a maioria dos conjuntos em todos os tipos de câncer, com algumas exceções para a ômica miRNA e um caso da ômica CNV (KIRP). O comportamento ob-

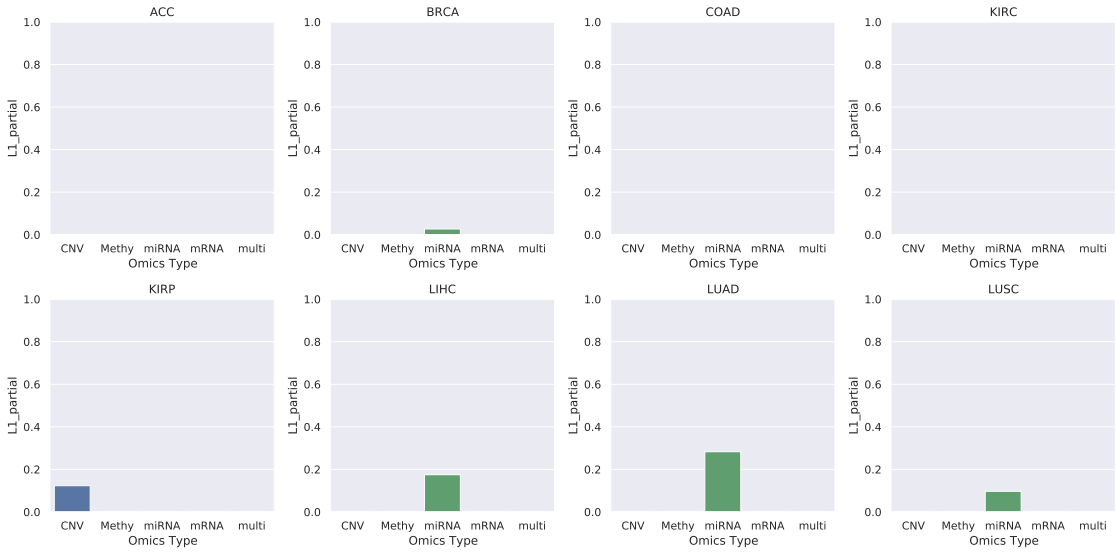
Figura 5.6 – Análise da medida de complexidade N3_Partial para a classe "No", por tipo de Câncer e tipo de Ômica



Fonte: O Autor

servado parece estar ligado à alta dimensionalidade de cada um dos conjuntos de dados ômicos, que em geral é bem maior que o número de instâncias. Assim, pode se imaginar que com uma alta dimensão, em algum ponto as classes se tornarão linearmente separáveis. Apoia esta suposição o fato de que o tipo de ômica onde ocorreram a maioria dos valores acima de zero é a que possui menor dimensionalidade (*i.e.,* miRNA).

Figura 5.7 – Análise da medida de complexidade L1_Partial para a classe "No", por tipo de Câncer tipo de Ômica



Fonte: O Autor

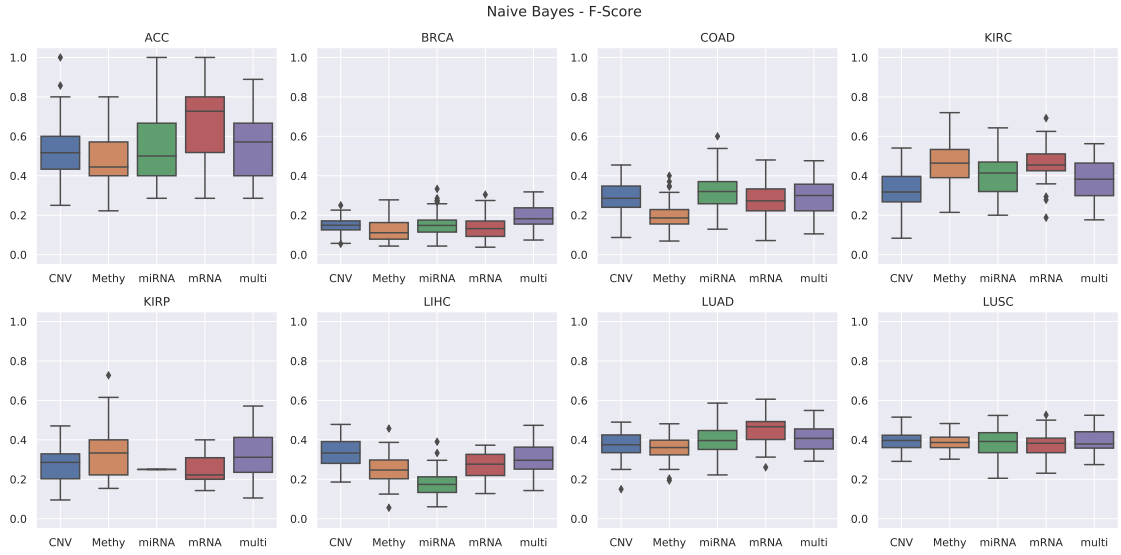
5.2 Desempenho dos Modelos na Tarefa de Classificação

Nesta seção, apresentamos os resultados das medidas de desempenho na tarefa de classificação obtidas para cada um dos dois modelos (*i.e.,* NB e GLM) treinados com cada conjunto de dados: CNV, Metilação de DNA, expressão de mRNA, expressão de miRNA e o conjunto multi-ômicas (*i.e.,* a concatenação de todos os anteriores). Com a utilização da estratégia de validação cruzada K-fold (com k=5) repetida 10 vezes, o número de estimativas de desempenho por modelo foi de 50 amostras. As figuras a seguir mostram os diagramas de caixa para os modelos NB e GLM das medidas F1-Score e Recall (a análise da precisão pode ser consultada no Apêndice A). Optamos por esta forma de apresentação devido à quantidade de medições realizadas em cada conjunto. Através do diagrama de caixas podemos observar informações relevantes, como a mediana e a dispersão dos resultados de desempenho. Notamos que os resultados do modelo GLM não incluem as estimativas para BRCA e KIRP, pois em ambos os casos o modelo apresentou dificuldades em lidar com o desbalanceamento entre classes e modelar adequadamente o problema, retornando sempre uma predição para a classe majoritária ("Yes").

As Figuras 5.8 e 5.9 mostram os valores de F1-Score para os modelos Naive Bayes e Regressão Logística, respectivamente. Podemos perceber que de uma forma geral as dispersões foram maiores para o modelo de Regressão Logística. Além disso, para ambos os modelos, o câncer ACC foi aquele que obteve desempenhos mais altos em termos de F1-Score, Recall e Precisão. Este achado é interessante, tendo em vista que em muitas medidas de complexidade os dados relacionados a ACC apresentavam valores menores, indicando menor complexidade para discriminação entre classes. Adicionalmente, observando a mediana de cada conjunto de dados, é possível notar que embora os níveis de desempenho para diferentes ômicas em um mesmo tipo de câncer possuam uma certa proximidade, de modo geral os resultados não possuem uma regularidade entre as ômicas. Isto é, enquanto para alguns tipos de tumores os melhores desempenhos podem ser observados com os dados de expressão de mRNA (por exemplo, ACC), para outros os desempenhos mais altos foram obtidos com metilação de DNA (por exemplo, KIRC). Também notamos que embora o conjunto de dados multi-ômicos seja o mais rico e completo em termos de informações sobre alterações moleculares em amostras de tumor, o seu uso não necessariamente reflete em melhor desempenho na classificação. De fato, apenas em casos pontuais ele foi capaz de efetivamente melhorar a predição de sobrevida dos pacientes. Este resultado também acompanha o que foi observado na análise de

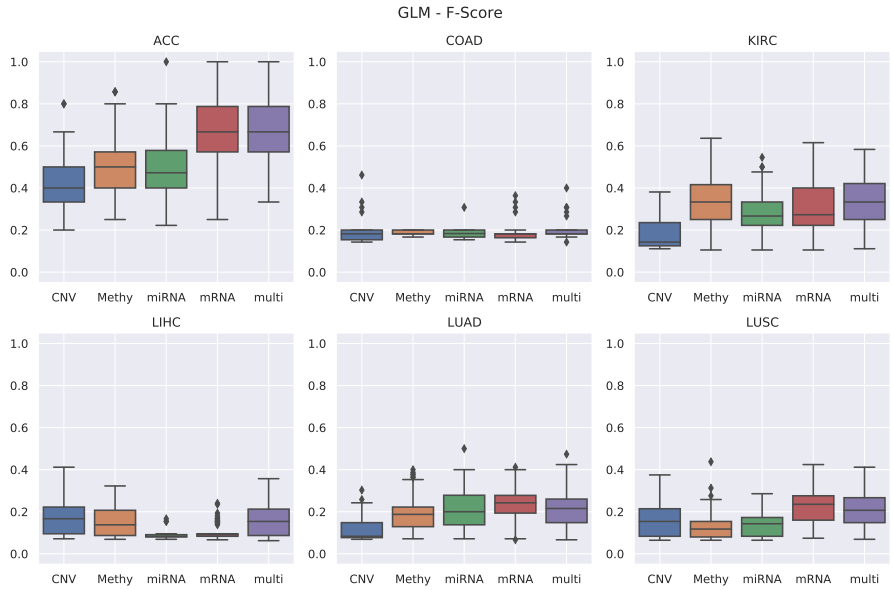
complexidade dos dados, visto que apenas em alguns casos o conjunto de dados multi-ômicos apresentou resultados notavelmente inferiores para as medidas de complexidade analisadas.

Figura 5.8 – Análise do F1-Score para o modelo Naive Bayes, por tipo de câncer e tipo de ômica



Fonte: O Autor

Figura 5.9 – Análise do F1-Score para o modelo de Regressão Logística, por tipo de câncer e tipo de ômica

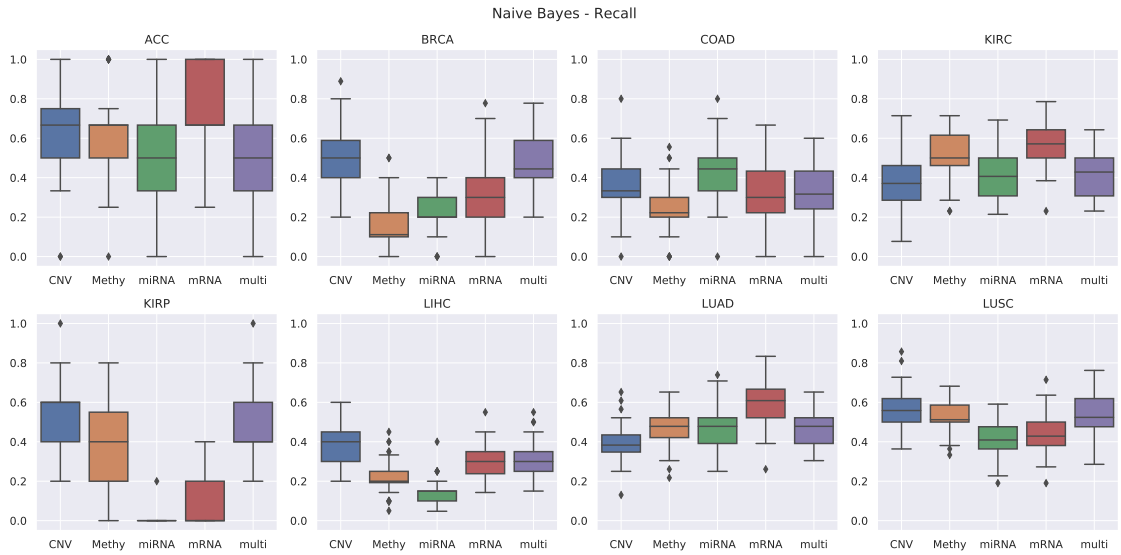


Fonte: O Autor

Ao analisar os valores de recall, apresentados nas Figuras 5.10 e 5.11, notamos com maior clareza que o modelo de Regressão Logística teve uma dificuldade maior na

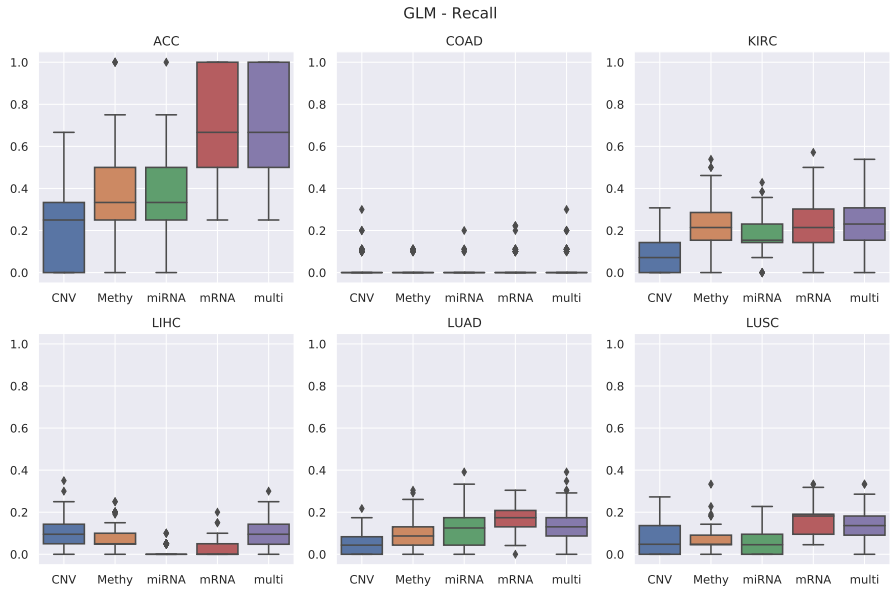
tarefa de classificação e especialmente na identificação das instâncias na classe de interesse ("No", indicando óbito em 3 anos após diagnóstico de tumor). Esta observação corrobora o fato de que embora as medidas de complexidade possam estimar o quão difícil é a tarefa de classificação de acordo com a natureza dos dados, na prática, os resultados sofrem influência do tipo de algoritmo de aprendizado aplicado, conforme esperado.

Figura 5.10 – Análise do recall para o modelo Naive Bayes, por tipo de câncer e tipo de ômica



Fonte: O Autor

Figura 5.11 – Análise do recall para o modelo de Regressão Logística, por tipo de câncer e tipo de ômica

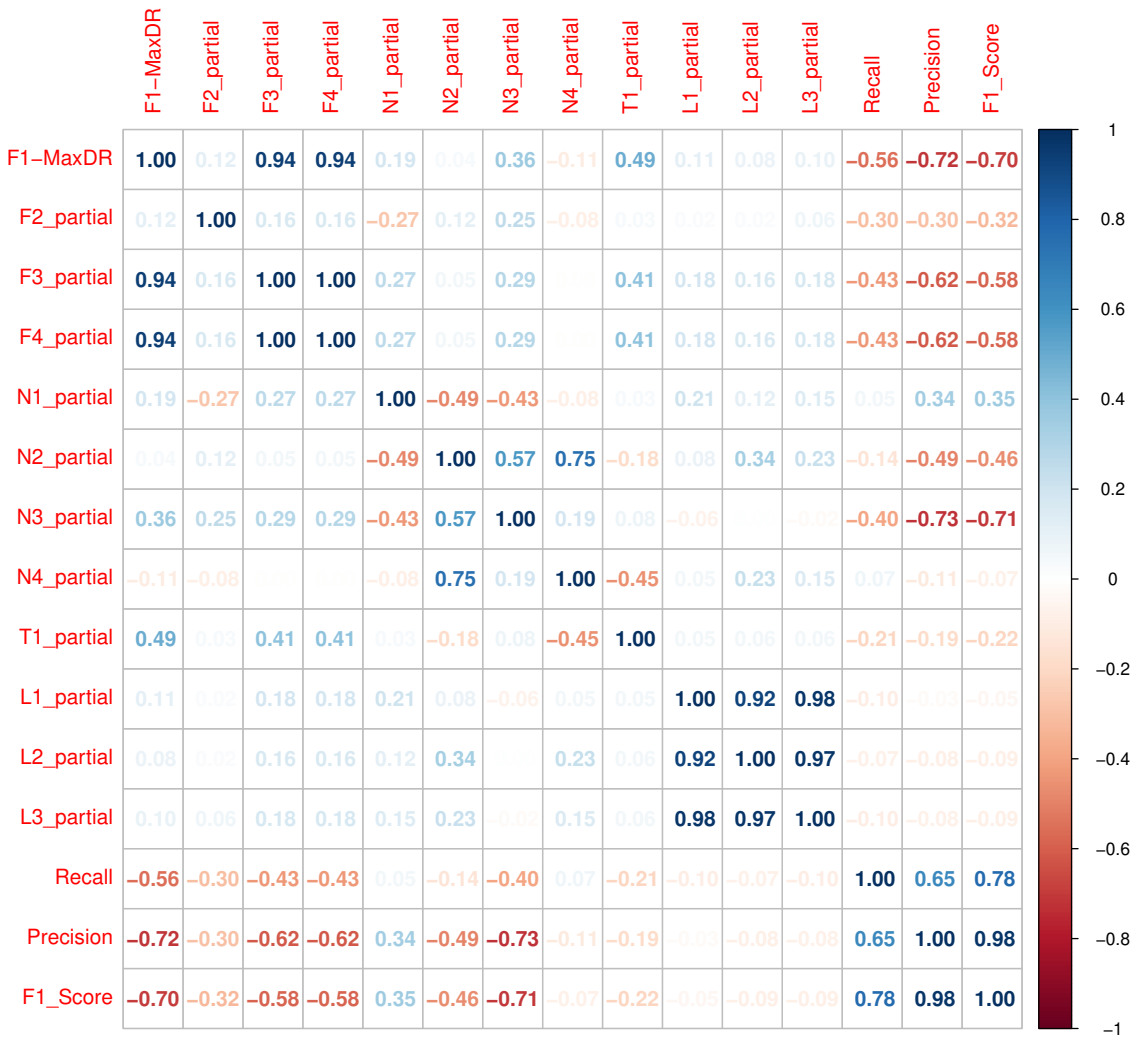


Fonte: O Autor

5.3 Correlações entre as medidas de complexidade e desempenho de modelos

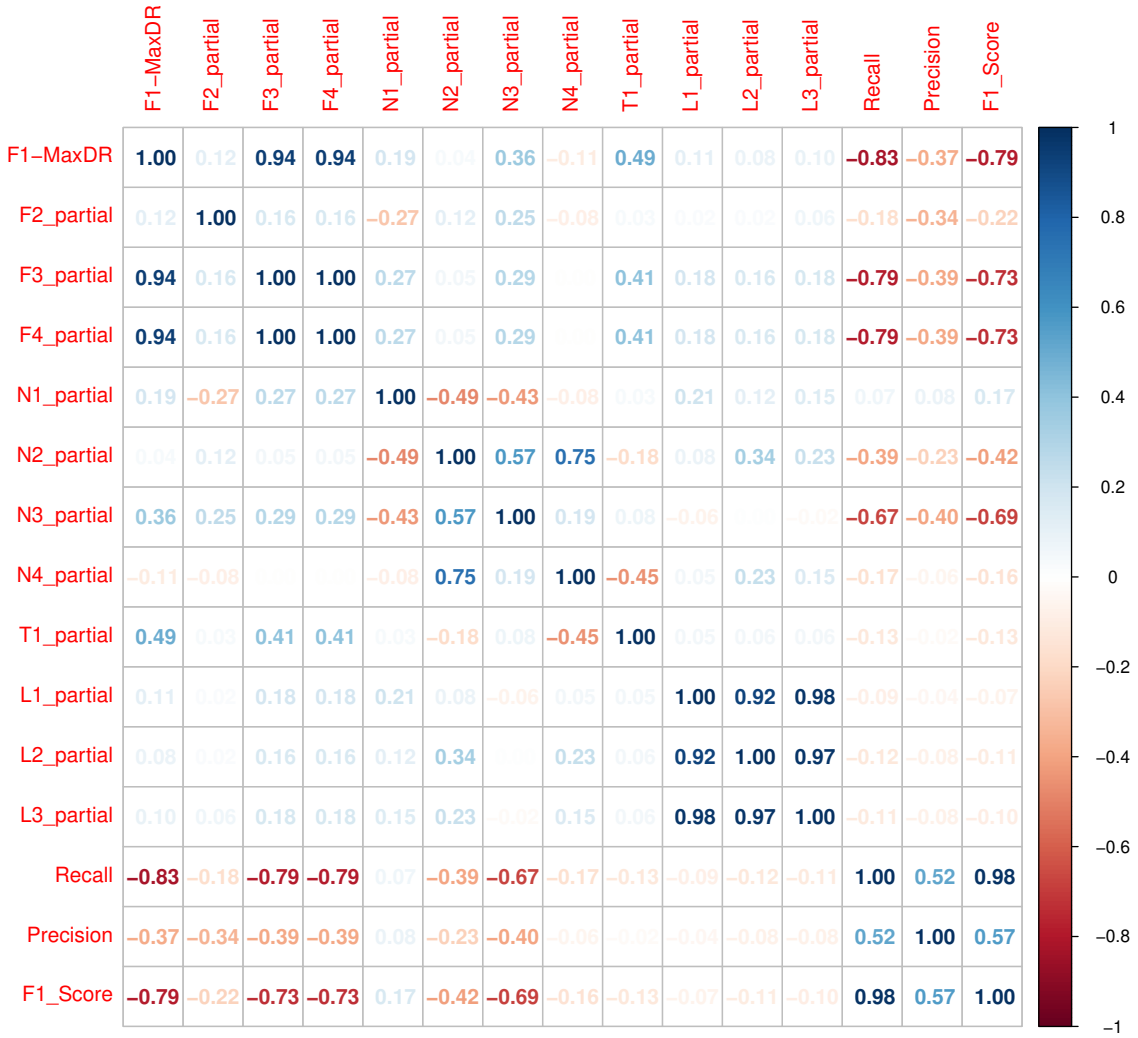
A fim de identificar correlações entre as medidas de complexidade e as métricas de desempenho, montamos duas tabelas, uma contendo todas as medidas de complexidade e as análises de desempenho do modelo Naive Bayes e outra contendo todas as medidas de complexidade e as análises de desempenho do modelo de Regressão Logística. A partir de cada uma destas tabelas, foi construída a matriz de correlação de Pearson entre as medidas de complexidade de dados e os desempenhos obtidos pelo modelo correspondente. As matrizes são apresentadas nas Figuras 5.12 e 5.13 para o Naive Bayes e a Regressão Logística, respectivamente.

Figura 5.12 – Correlação entre as medidas de complexidade de dados e as métricas de desempenho do modelo Naive Bayes



Fonte: O Autor

Figura 5.13 – Correlação entre as medidas de complexidade de dados e as métricas de desempenho do modelo de Regressão Logística



Fonte: O Autor

Conforme é possível observar, quase todas as correlações entre as medidas de complexidade de dados e as métricas de desempenho são negativas. Este comportamento é esperado, uma vez que se espera um menor desempenho (métricas de desempenho próximas de 0) em tarefas mais complexas (medidas de complexidade próximas de 1). Além disso, os dois modelos apresentaram, de modo geral, comportamento semelhante na correlação entre os dois grupos de medidas, complexidade e desempenho. Isso nos permite afirmar que embora os valores absolutos de desempenho em termos de F1, Recall ou Precisão possam diferir entre os modelos, as tendências de associação de desempenho preditivo com a complexidade da tarefa de predição tendem a ser similares entre os mesmos. Em números absolutos, F1-MaxDR, F3_Partial, F4_Partial e N3_Partial apresentaram, para ambos os modelos, correlações acima de 0.5 com pelo menos uma das métricas de

desempenho.

Entre as medidas de complexidade, F1-MaxDR foi a que apresentou as correlações mais altas com as métricas de desempenho, chegando a -0.83 para o recall no modelo de Regressão Logística (Figura 5.13). N3_Partial obteve correlações com o desempenho bastante parecidas com as que obteve F1-MaxDR, porém levemente menores em valores absolutos quando considerada a média dos dois modelos. As correlações que F3_Partial e F4_Partial apresentaram com as métricas de desempenho foram exatamente as mesmas, com resultados bastante parecidos também com N3_Partial. F2_Partial e N2_Partial apresentaram correlações com valor absoluto abaixo de 0.50, mas de um modo geral acima de 0.20. Também chama a atenção o fato de N1_Partial possuir uma correlação positiva com as medidas de desempenho. Para o modelo de Regressão Logística, estes valores são bem pouco significativos, mas, no modelo NB, apesar de baixas, as correlações são mais relevantes.

A respeito das medidas de desempenho, as correlações para o Recall do modelo de Regressão Logística foram no geral maiores que para o modelo NB. Já a Precisão apresentou comportamento inverso, com correlações mais fortes em NB e mais fracas na Regressão Logística. F1-score apresentou correlações mais fortes para o modelo de Regressão Logística, entretanto, em ambos os modelos os resultados ficaram mais próximos do maior valor absoluto apresentado entre Precisão e Recall.

Fazendo uma análise dos resultados de correlação, nossas observações mostraram uma correlação forte entre as medidas F1-MaxDR, F3_Partial, F4_Partial e N3_Partial, e o desempenho dos modelos de classificação. Em algumas situações, estas medidas podem funcionar como preditores do nível de desempenho. Nesse sentido, uma complexidade menor pode indicar a tendência de um desempenho mais elevado.

Sobre F2_Partial e N2_Partial, estas apresentam uma correlação moderadamente fraca com as medidas de desempenho, mas não tão próximas de zero. Parece haver a possibilidade de que as medidas possam ter maior ou menor utilidade em condições distintas. Talvez seja interessante avaliar futuramente as correlações destas medidas para outros tipos de classificadores. No caso de F2_Partial, ainda há questão apontada na Seção 5.1, a respeito da variação na escala de valores da medida.

É importante notar que a correlação permite saber se há uma relação linear entre as medidas de complexidade e o desempenho, mas não é possível determinar os valores específicos de desempenho ou a proporção exata de aumento ou decréscimo deste. Além disso, é necessário o uso de um ou mais conjuntos como base de comparação, para de-

terminar uma referência em relação à tendência de aumento ou redução do desempenho. Nesse caso a escolha da base de comparação pode influenciar na qualidade das predições. É importante que os conjuntos de referência possuam alguma semelhança com os conjuntos avaliados, compartilhando características que permitam a comparação. Uma possível aplicação seria avaliar se determinada técnica de pré-processamento reduz a complexidade dos dados e, portanto, tende a melhorar o desempenho do modelo final. Ou ainda comparar diversos conjuntos de dados do mesmo campo de conhecimento, em busca de certas regularidades. Neste trabalho, usamos esse tipo de comparação para avaliar se há semelhanças entre os tipos de câncer, e também entre os tipos de dados ômicos usados.

Também é recomendado que para fazer eventuais análises em direção a prever a dificuldade de uma tarefa de predição, se utilize mais de uma das medidas de complexidade de dados como preditor, de forma que as conclusões sejam mais robustas. Como evidência, tomemos a correlação de F1_Score com F3_Partial e N3_Partial. No modelo NB a correlação de F3_Partial (-0.58) é mais fraca que a de N3_Partial (-0.71), e a situação se inverte no modelo de Regressão Logística (-0.73 e -0.69, respectivamente). A escolha da medida de desempenho também afeta os resultados, como pode ser observado pela mudança de comportamento nas correlações das medidas de precisão e recall em cada um dos modelos. Acreditamos que a análise conjunta das medidas forneça informações melhores do que o uso de uma única medida como indicativo da melhora de desempenho.

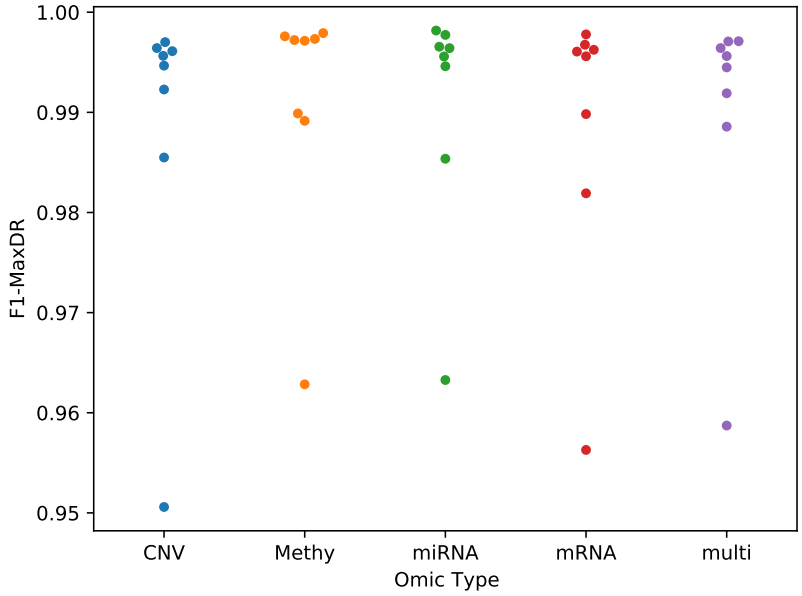
5.4 Comparação Entre Tipos de Dados Ômicos

Nesta seção, tentaremos responder às perguntas sobre a semelhança entre dados de diferentes ômicas e possíveis vantagens ou desvantagens da abordagem multi-ômica. Para tanto, utilizamos uma série de *swarmplots* que mostram a distribuição dos valores de complexidade de cada tipo de câncer, dada uma medida de complexidade e um tipo de dado ômico. Por exemplo, cada ponto no *swarmplot* da Figura 5.14 corresponde a um tipo de câncer, enquanto o eixo x indica o tipo de ômica e o eixo y indica a medida de complexidade analisada. Para cada tipo de ômica, há 8 pontos, que correspondem aos 8 tipos de câncer estudados.

Do ponto de vista da complexidade, conforme comentado na Seção 5.1, o comportamento dos diferentes tipos ômicos é bastante semelhante para a maioria das medidas que possuem maior correlação com o desempenho, dado um tipo de câncer. Utilizando as medidas com altas correlações com o desempenho preditivo (Seção 5.3), represen-

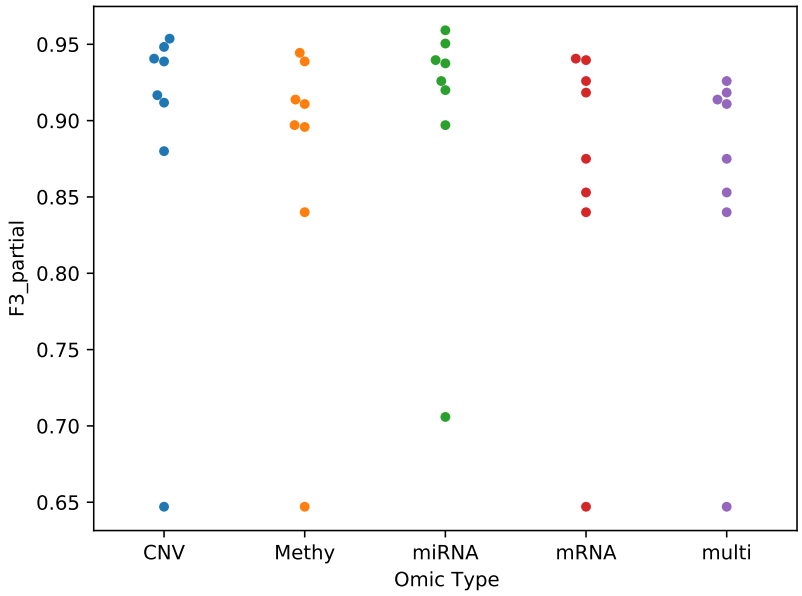
tando potenciais preditores de complexidade de classificação, é possível identificar a semelhança no comportamento. As Figuras 5.14, 5.15 e 5.16 apresentam o *swarmplot* das ômicas para a medida correspondente: F1-MaxDR, F3_Partial e N3_Partial. Nota-se que o intervalo de valores ocupados em cada ômica é muito semelhante aos das outras.

Figura 5.14 – Swarmplot da distribuição de F1-MaxDR para as diferentes ômicas



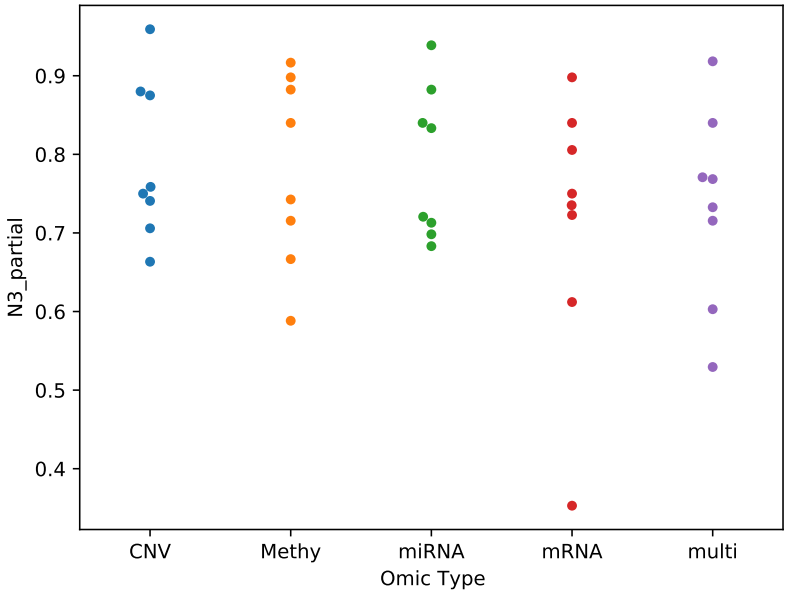
Fonte: O Autor

Figura 5.15 – Swarmplot da distribuição de F3_Partial para as diferentes ômicas



Fonte: O Autor

Figura 5.16 – Swarmplot da distribuição de N3_Partial para as diferentes ômicas



Fonte: O Autor

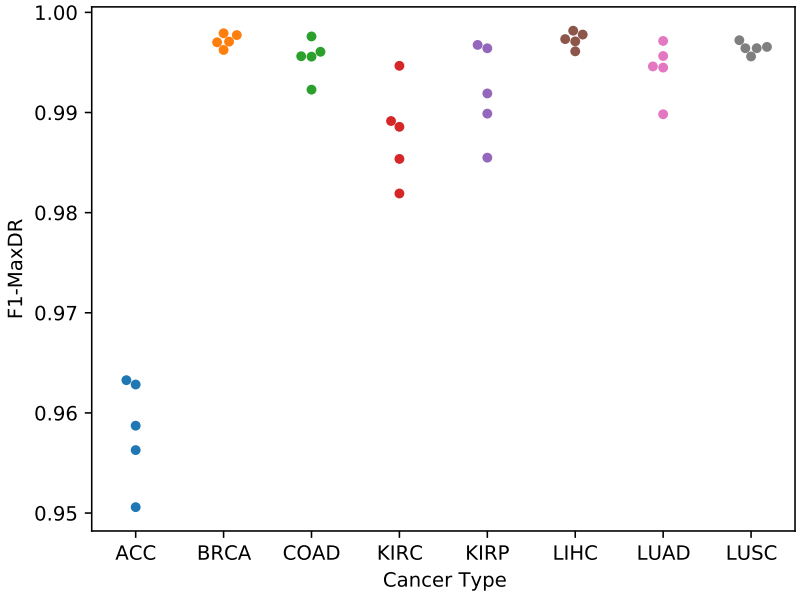
Levando-se em conta o desempenho dos modelos, é possível perceber que as semelhanças no comportamento não são tão acentuadas, mas estão presentes. Observando novamente as Figuras 5.8 e 5.9, pode-se perceber que o intervalo interquartil compartilha faixas de valores entre as ômicas, para a maioria dos tipos de câncer. Comparando os gráficos da medidas de complexidade e desempenho, se pode notar, em certos casos de forma mais fácil, que há certa relação entre os comportamentos. No entanto, através das medidas selecionadas de complexidade, não conseguimos identificar uma ômica que se destaque particularmente. Ou seja, não pudemos apontar um dos tipos que seja significativamente mais fácil ou mais difícil para todos os tipos de câncer. O mesmo pode ser dito analisando as medidas de desempenho para os modelos treinados.

Nossa suposição inicial de que o conjunto de dados multi-ômico poderia ter melhores resultados do que os tipos ômicos individuais na predição de sobrevida também não pôde ser comprovada com os resultados gerados. No entanto, o tipo multi-ômico nunca tem o resultado abaixo de todas as outras ômicas. Adicionalmente, frequentemente dentro de um tipo de câncer, os dados multi-ômicos possuem resultados bastante próximos ou iguais àqueles atingidos pela melhor ômica. Desta forma, pode-se sugerir que o uso de dados multi-ômicos introduz certa robustez na análise preditiva.

5.5 Comparação entre tipos de Câncer

Realizamos também análises a respeito do comportamento dos tipos de câncer para as medidas de complexidade analisadas. Com esta comparação, procuramos estabelecer se há tipos semelhantes entre si, ou algum que seja menos complexo. De maneira análoga ao estudo para os dados ômicos, utilizamos uma série de *swarmplots* que mostram a distribuição dos valores de complexidade de cada tipo de ômica, dada uma medida de complexidade e um tipo de câncer. Apresentamos a seguir as medidas F1-MaxDR (Figura 5.17), F3_Partial (Figura 5.18) e N3_Partial (Figura 5.19); a comparação para as demais medidas de complexidade pode ser consultada no Apêndice A.

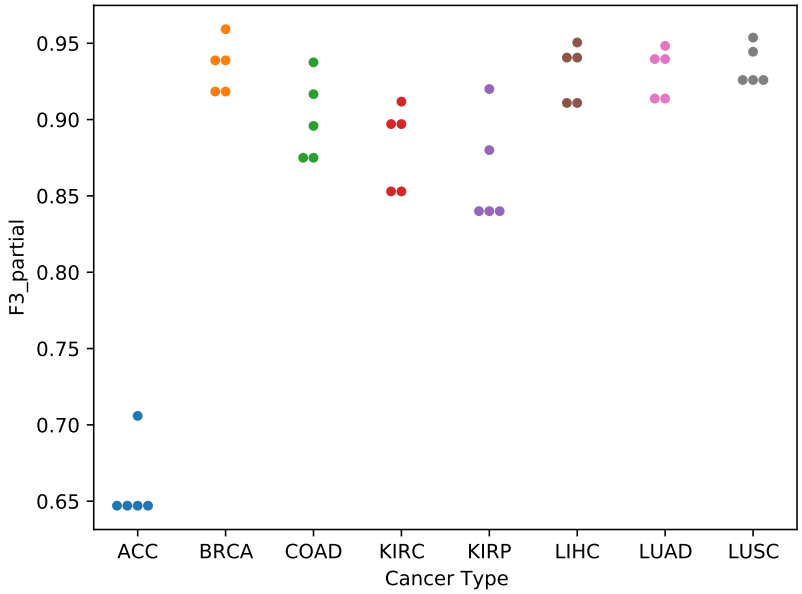
Figura 5.17 – Swarmplot da distribuição de F1_MaxDR para os tipos de câncer



Fonte: O Autor

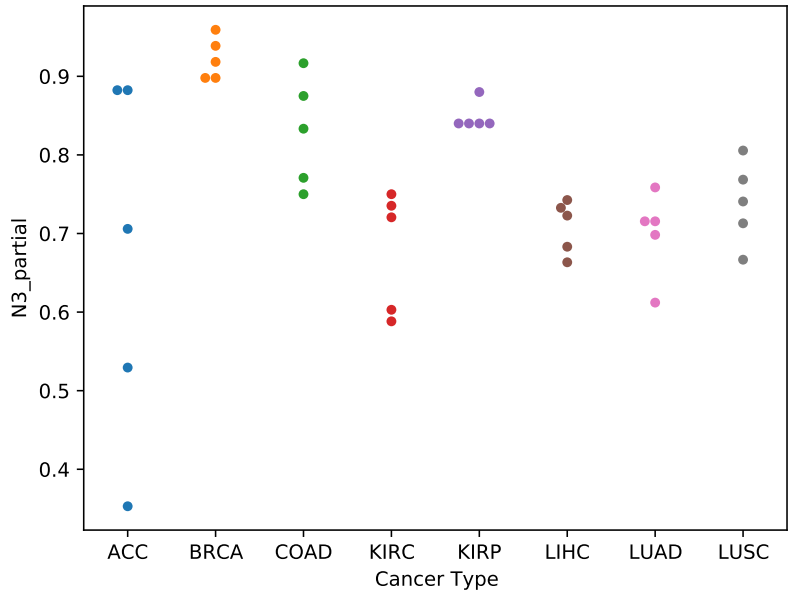
Conforme apontado na Seção 5.1, na análise das medidas de complexidade tivemos como destaque o câncer ACC. Os conjuntos de dados para este tipo de tumor apresentaram, de forma consistente, uma complexidade menor segundo as medidas de complexidades mais correlacionadas com as medidas de desempenho em nossas observações. Esta diferença fica clara nas Figuras 5.17 e 5.18, que representam medidas de complexidade que analisam a sobreposição de *features*. Os resultados da análise de desempenho corroboram esta percepção. Os valores de desempenho do tipo de câncer ACC são consistentemente melhores, nos dois modelos. Esta é uma evidência que apoia a ideia de que as medidas de complexidade podem ser usadas para predizer a tendência

Figura 5.18 – Swarmplot da distribuição de F3_Partial para os tipos de câncer



Fonte: O Autor

Figura 5.19 – Swarmplot da distribuição de N3_Partial para os tipos de câncer



Fonte: O Autor

de melhora no desempenho. Para a medida N3_Partial (Figure 5.19), embora algumas ômicas para o ACC possuam valores no intervalo [0.7, 1.0], onde a maioria dos demais tipos de câncer varia para esta medida, o ACC possui dois tipos de ômicas com valores de N3_Partial abaixo de 0.6, sendo os menores valores para esta medida dentre todos os

conjuntos de dados analisados.

Quanto aos demais tipos de câncer, de um modo geral eles apresentam mais semelhanças quando comparados entre si com relação à complexidade do que com relação ao desempenho, embora ainda seja possível notar similaridades no segundo caso. A diferença entre os níveis de semelhança para complexidade e desempenho faz sentido e, de certo modo, é esperada. A correlação de Pearson permite estabelecer a força de uma relação linear, mas não determinar a função exata desta relação. Quanto mais forte a correlação, maior a semelhança entre o comportamentos dos dois tipos de medida em termos de tendência; isto é, quando um aumenta (diminui) o outro diminui (aumenta), no caso de existirem correlações. Entretanto, a correlação não nos permite aferir sobre os valores absolutos de desempenho, os quais podem variar entre dois modelos ainda que ambos demonstrem correlação perfeita entre medidas de complexidade e métricas de desempenho.

6 CONCLUSÃO

As grandes quantidades de dados ômicos gerados atualmente fornecem um vasto campo para a aplicação de técnicas de AM no tratamento e transformação destas informações em conhecimento. Certas características inerentes desta área aumentam a complexidade dos dados, como sobreposição entre as classes, e fronteiras de decisão mais complexas. Estes são fatores comumente associados a um menor desempenho dos modelos de aprendizado de máquina. Outra característica importante, frequentemente presente em dados ômicos, é o desbalanceamento de classes, que pode alterar o resultado de avaliações da complexidade e o seu efeito sobre o desempenho. Nesse sentido, é essencial o uso de boas medidas de complexidade, adaptadas para o contexto de dados desbalanceados.

O presente trabalho buscou fazer o uso dessas medidas de complexidades adaptadas para domínios com classes desbalanceadas a fim de verificar se é possível estabelecer uma conexão entre as características intrínsecas dos dados capturadas por estas medidas e as métricas de desempenho para uma série de dados ômicos ligados ao diagnóstico clínico de câncer. Tentamos também compreender as possíveis relações existentes entre os tipos de dados ômicos para diferentes tipos de câncer, e eventuais semelhanças ou diferenças entre os tipos de câncer. A partir das observações realizadas foi possível estabelecer que, dentro de um contexto de dados desbalanceados, há medidas de complexidade com forte correlação com as medidas de desempenho adotadas para a classe minoritária. Notadamente F1-MaxDR, F3_Partial, F4_Partial e N3_Partial, uma vez estabelecida uma base de comparação, podem servir como indicadores da tendência do aumento de desempenho em classificadores. Muito embora não seja possível quantificar a proporção exata do aumento de desempenho, uma indicação da tendência pode ajudar em diversas situações. Podemos citar como exemplos, a escolha e ajuste dos modelos, escolha de métodos de pré-processamento dos dados, comparação entre conjuntos de dados de um mesmo domínio, entre outros (LORENA et al., 2019).

Também realizamos a comparação entre diferentes tipos de ômicas e diferentes tipos de câncer, utilizando medidas de complexidade que selecionamos a partir do passo anterior de análise de correlações. Pudemos identificar um tipo de câncer que possui complexidade significativamente menor em relação aos outros (ACC) - um achado que vale a pena ser melhor explorado em trabalhos futuros. Encontramos, ainda, uma semelhança nos níveis de complexidade entre os diferentes tipos de ômicas. As análises de desempenho obtidas durante o treinamento dos modelos mostraram que o tipo de câncer ACC

obteve resultados melhores que os demais tipos de câncer. Para os diferentes tipos de ômicas, os níveis de desempenho dos modelos seguiram aproximadamente as tendências indicadas por algumas das medidas.

Embora tenham sido encontrados resultados interessantes, ainda permanecem muitas questões não respondidas, ou que poderiam ser melhor investigadas em trabalhos futuros. Este trabalho realizou o estudo considerando um conjunto restrito de medidas de desempenho e de algoritmos de classificação. A realização de estudos adicionais envolvendo outros algoritmos de classificação e a inclusão de outras medidas de desempenho podem ajudar a compreender melhor a relação entre complexidade e desempenho. Da mesma forma, poderiam ser feitas pesquisas sobre algumas das medidas de complexidade, como F2_Partial, buscando por adaptações que possam resolver as limitações encontradas. Também é interessante expandir este estudo para diferentes tarefas preditivas relacionadas ao diagnóstico e prognóstico de câncer a fim de verificar se os achados se repetem entre diferentes tipos de tarefas de classificação, possibilitando a extração de conclusões mais amplas sobre a utilidade de determinados dados ômicos para auxiliar na tomada de decisão clínica. Estas são questões que deveriam ser abordadas em outros trabalhos, dada a sua relevância quando se fala em tarefas de classificação, em particular no domínio de dados ômicos aplicados ao estudo de doenças complexas.

REFERÊNCIAS

BARELLA, V. H. et al. Assessing the data complexity of imbalanced datasets. **Information Sciences**, Elsevier, v. 553, p. 83–109, 2021.

BARELLA, V. H. et al. Data complexity measures for imbalanced classification tasks. In: IEEE. **2018 International Joint Conference on Neural Networks (IJCNN)**. [S.l.], 2018. p. 1–8.

BOLÓN-CANEDO, V.; MORAN-FERNANDEZ, L.; ALONSO-BETANZOS, A. An insight on complexity measures and classification in microarray data. In: IEEE. **2015 International Joint Conference on Neural Networks (IJCNN)**. [S.l.], 2015. p. 1–8.

CAI, Z. et al. Machine learning for multi-omics data integration in cancer. **iScience**, v. 25, n. 2, p. 103798, 2022. ISSN 2589-0042.

DUAN, R. et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. **PLOS Computational Biology**, Public Library of Science, v. 17, n. 8, p. 1–33, 08 2021.

HO, T. K. A data complexity analysis of comparative advantages of decision forest constructors. **Pattern Analysis & Applications**, Springer, v. 5, n. 2, p. 102–112, 2002.

HO, T. K.; BASU, M. Complexity measures of supervised classification problems. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 3, p. 289–300, 2002. ISSN 01628828.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, p. 1–26, 2008.

LI, J. et al. Predicting breast cancer 5-year survival using machine learning: A systematic review. **PLOS ONE**, Public Library of Science, v. 16, n. 4, p. 1–23, 04 2021.

LIÑARES-BLANCO, J.; PAZOS, A.; FERNANDEZ-LOZANO, C. Machine learning analysis of TCGA cancer data. **PeerJ Computer Science**, PeerJ Inc., v. 7, p. e584, 2021.

LORENA, A. C. et al. Analysis of complexity indices for classification problems: Cancer gene expression data. **Neurocomputing**, Elsevier, v. 75, n. 1, p. 33–42, 2012.

LORENA, A. C. et al. **Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição)**. [S.l.]: LTC, 2011. ISBN 9788521637493.

LORENA, A. C. et al. How complex is your classification problem? a survey on measuring classification complexity. **ACM Computing Surveys**, ACM New York, NY, USA, v. 52, n. 5, p. 1–34, 2019.

LORENA, A. C. et al. On the complexity of gene marker selection. In: **2010 Eleventh Brazilian Symposium on Neural Networks**. [S.l.: s.n.], 2010. p. 85–90.

MORÁN-FERNÁNDEZ, L.; BOLÓN-CANEDO, V.; ALONSO-BETANZOS, A. Can classification performance be predicted by complexity measures? a study using microarray data. **Knowledge and Information Systems**, Springer, v. 51, n. 3, p. 1067–1090, 2017.

OKUN, O.; PRIISALU, H. Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors. **Artificial Intelligence in Medicine**, Elsevier, v. 45, n. 2-3, p. 151–162, 2009.

SÁNCHEZ, J. S.; GARCÍA, V. Addressing the links between dimensionality and data characteristics in gene-expression microarrays. In: **Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications**. [S.l.: s.n.], 2018. p. 1–6.

SOUTO, M. C. P. de et al. Complexity measures of supervised classifications tasks: A case study for cancer gene expression data. In: **The 2010 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2010. p. 1–7.

TONG, L. et al. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. **BMC Medical Informatics and Decision Making**, Springer, v. 20, n. 1, p. 1–12, 2020.

ZHAO, D. et al. Pan-cancer survival classification with clinicopathological and targeted gene expression features. **Cancer Informatics**, v. 20, p. 11769351211035137, 2021. PMID: 34376966.

APÊNDICE A — FIGURAS E GRÁFICOS ADICIONAIS

Neste apêndice, estão as figuras e gráficos adicionais dos experimentos realizados.

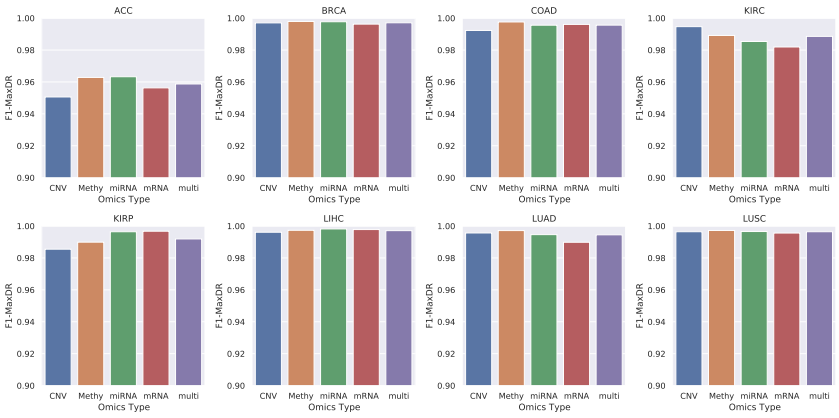
A.1 Comparação entre Ômicas por Tipo de Câncer

As figuras a seguir comparam os tipos de dados ômicos lado a lados, agrupados por tipos de câncer.

A.1.1 Medidas de Sobreposição de Atributos

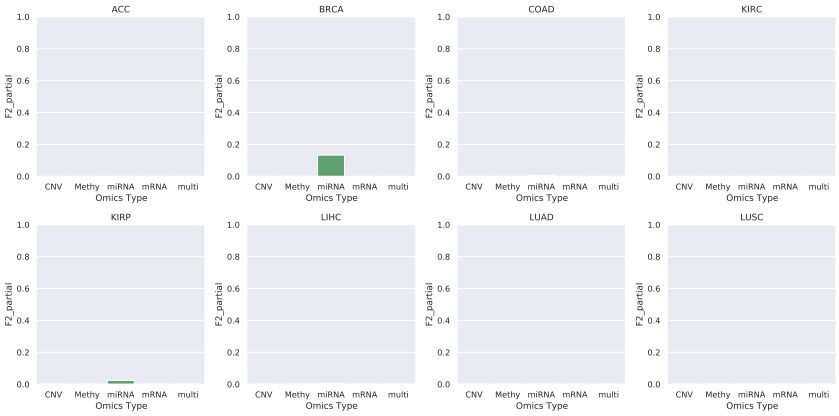
As Figuras A.1 a A.7 contém os resultados para as medidas de sobreposição de atributos.

Figura A.1 – F1-MaxDR - Ômicas por tipo de Câncer



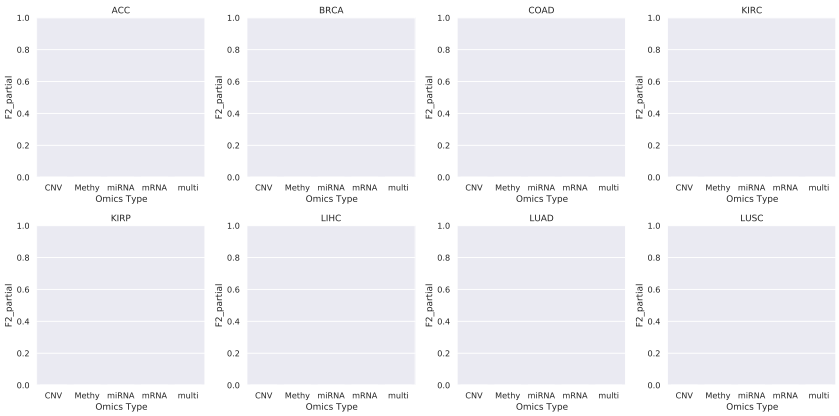
Fonte: O Autor

Figura A.2 – F2_Partial - Classe No - Ômicas por tipo de Câncer



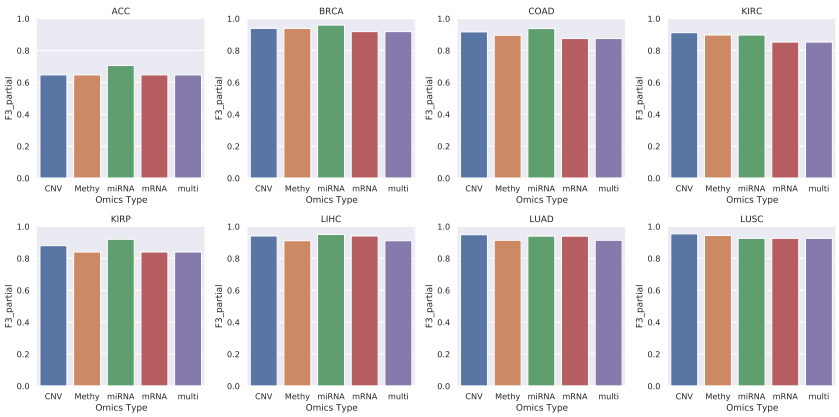
Fonte: O Autor

Figura A.3 – F2_Partial - Classe Yes - Ômicas por tipo de Câncer



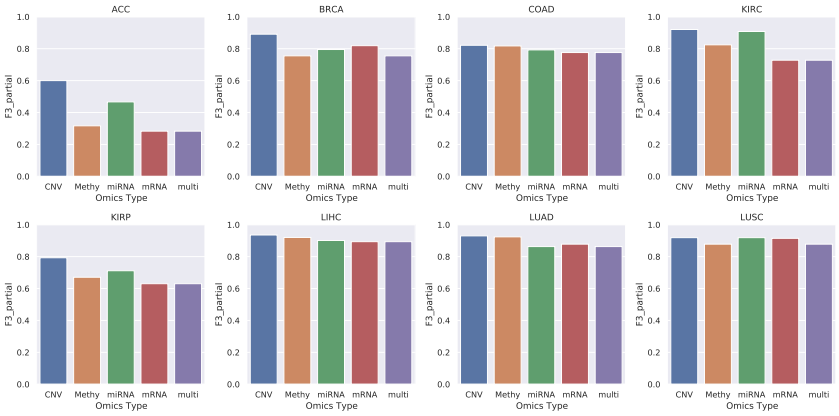
Fonte: O Autor

Figura A.4 – F3_Partial - Classe No - Ômicas por tipo de Câncer



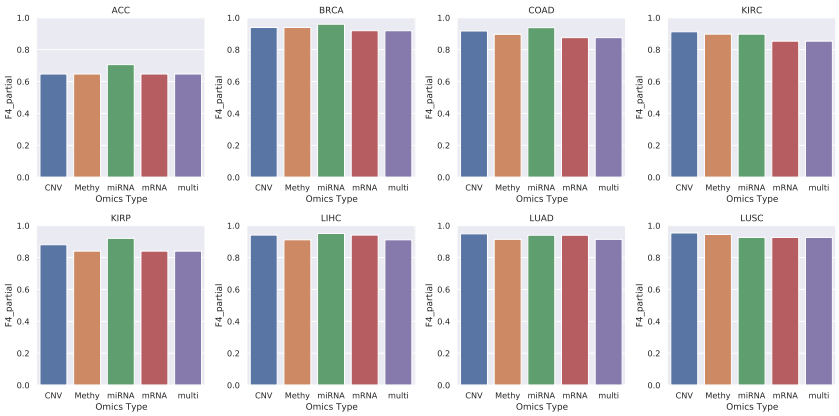
Fonte: O Autor

Figura A.5 – F3_Partial - Classe Yes - Ômicas por tipo de Câncer



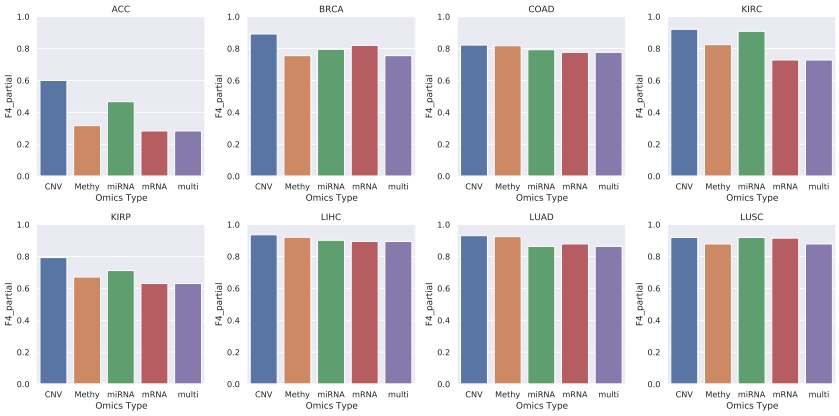
Fonte: O Autor

Figura A.6 – F4_Partial - Classe No - Ômicas por tipo de Câncer



Fonte: O Autor

Figura A.7 – F4_Partial - Classe Yes - Ômicas por tipo de Câncer

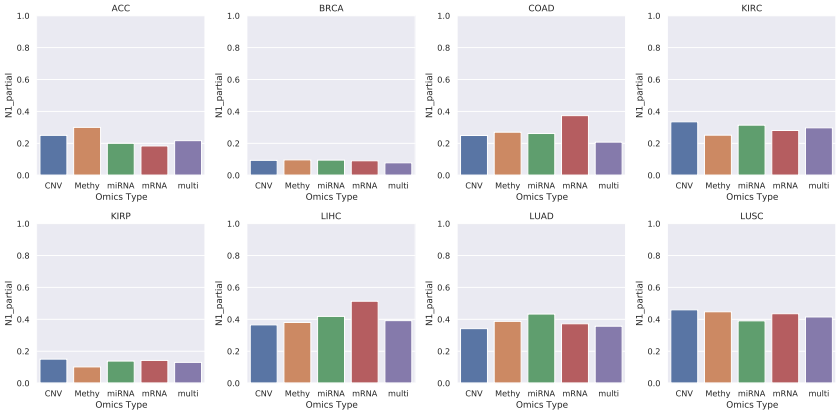


Fonte: O Autor

A.1.2 Medidas de Vizinhança

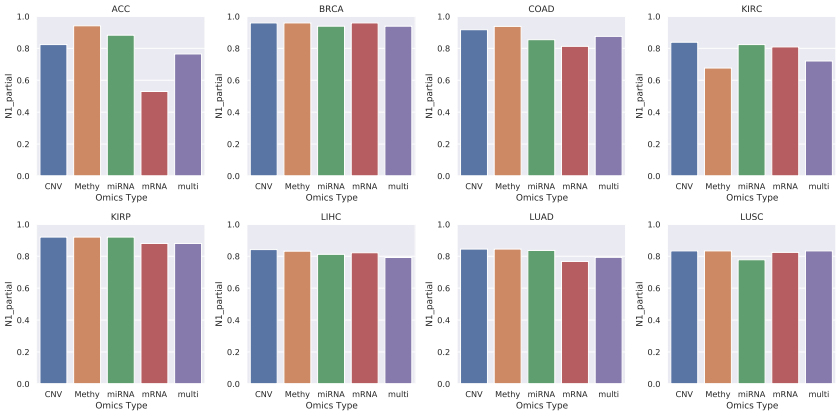
Esta seção contém os resultados para as medidas baseadas na vizinhança das instâncias, expressos na forma das Figuras A.8 a A.17.

Figura A.8 – N1_Partial - Classe No - Ômicas por tipo de Câncer



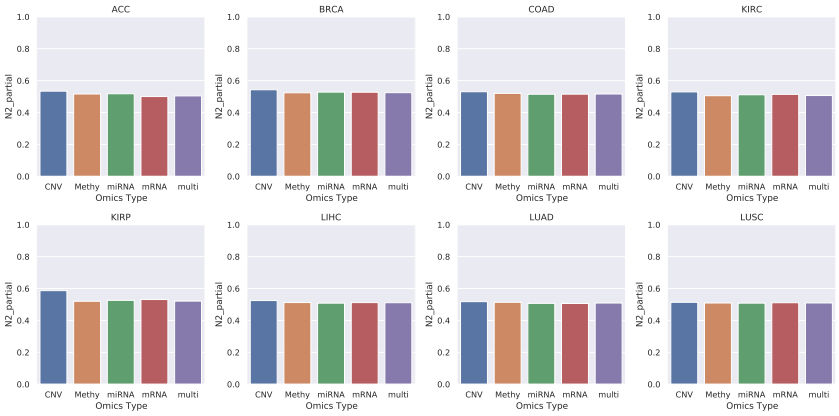
Fonte: O Autor

Figura A.9 – N1_Partial - Classe Yes - Ômicas por tipo de Câncer



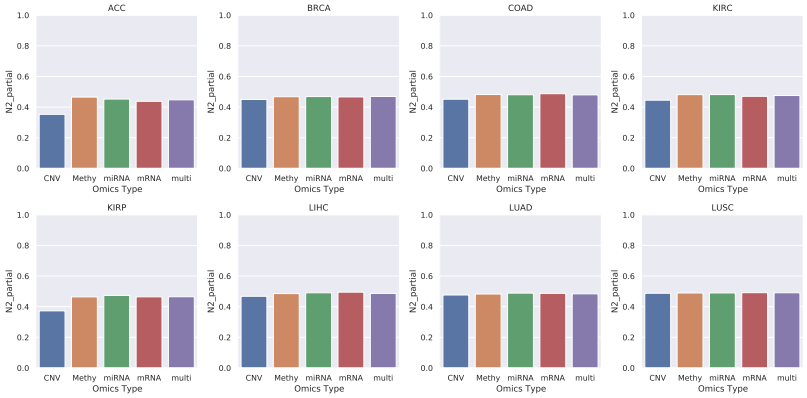
Fonte: O Autor

Figura A.10 – N2_Partial - Classe No - Ômicas por tipo de Câncer



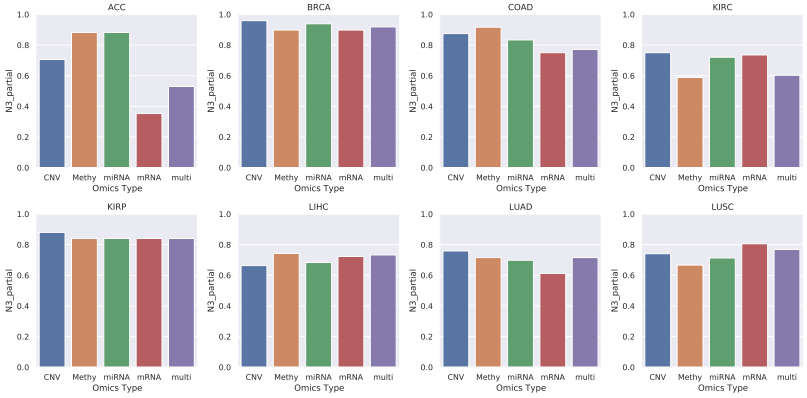
Fonte: O Autor

Figura A.11 – N2_Partial - Classe Yes - Ômicas por tipo de Câncer



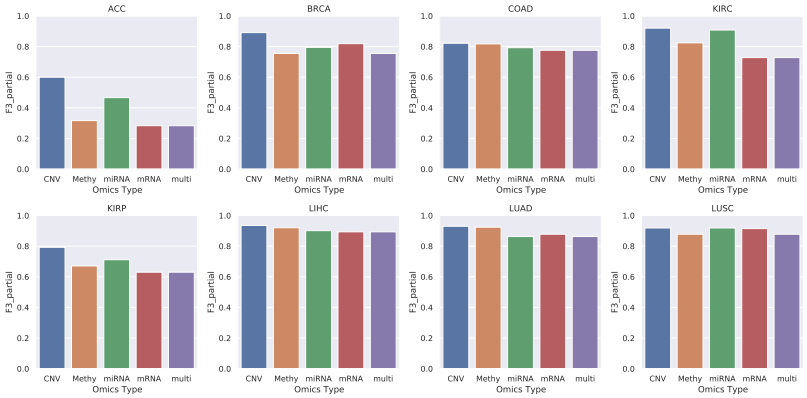
Fonte: O Autor

Figura A.12 – N3_Partial - Classe No - Ômicas por tipo de Câncer



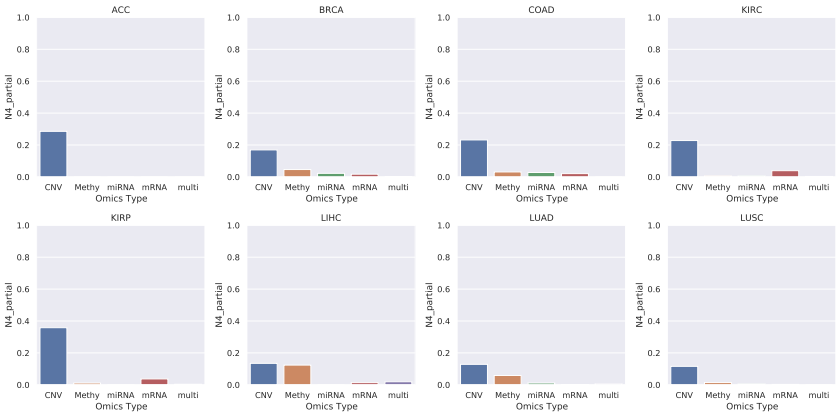
Fonte: O Autor

Figura A.13 – N3_Partial - Classe Yes - Ômicas por tipo de Câncer



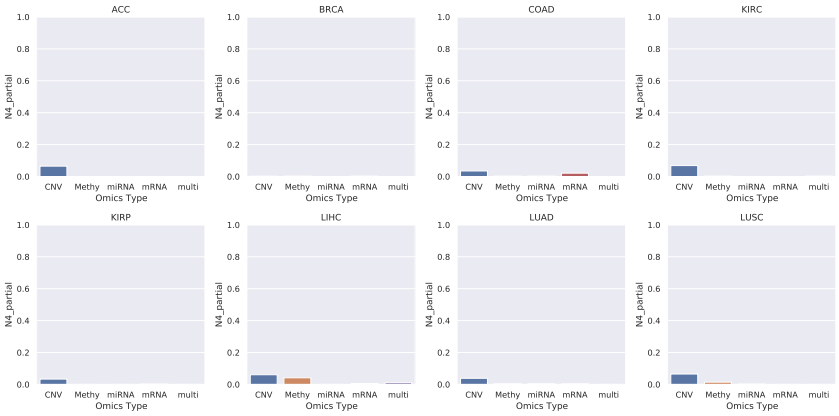
Fonte: O Autor

Figura A.14 – N4_Partial - Classe No - Ômicas por tipo de Câncer



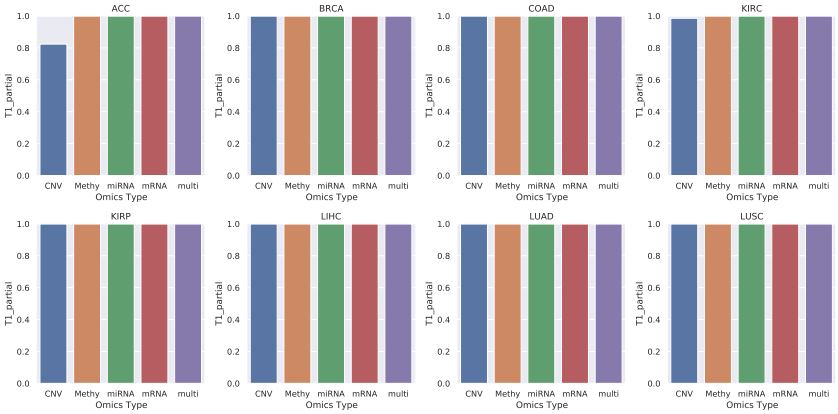
Fonte: O Autor

Figura A.15 – N4_Partial - Classe Yes - Ômicas por tipo de Câncer



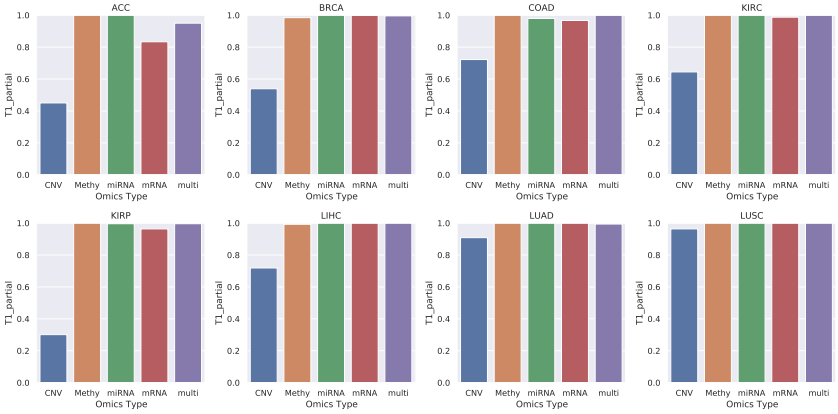
Fonte: O Autor

Figura A.16 – T1_Partial - Classe No - Ômicas por tipo de Câncer



Fonte: O Autor

Figura A.17 – T1_Partial - Classe Yes - Ômicas por tipo de Câncer

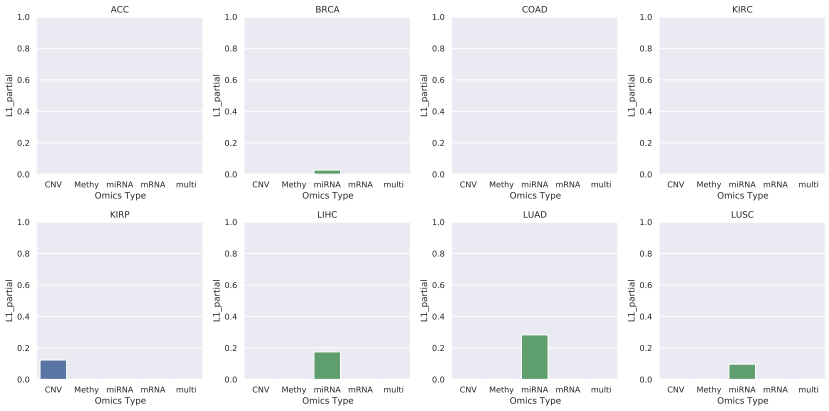


Fonte: O Autor

A.1.3 Medidas de Linearidade

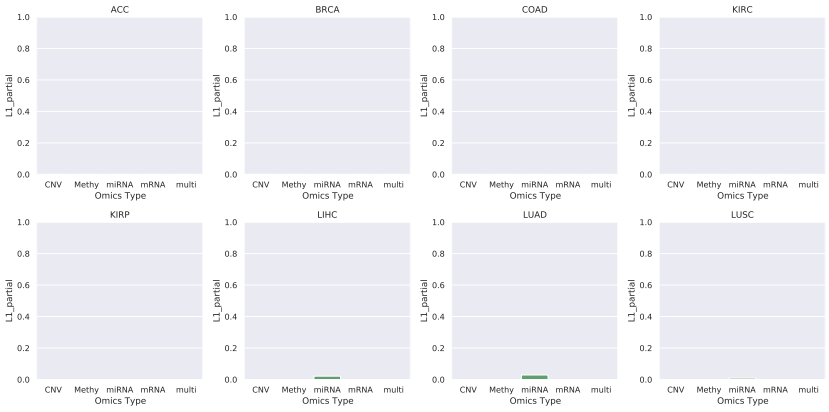
As Figuras A.18 a A.23 são relativas às métricas de separabilidade linear.

Figura A.18 – L1_Partial - Classe No - Ômicas por tipo de Câncer



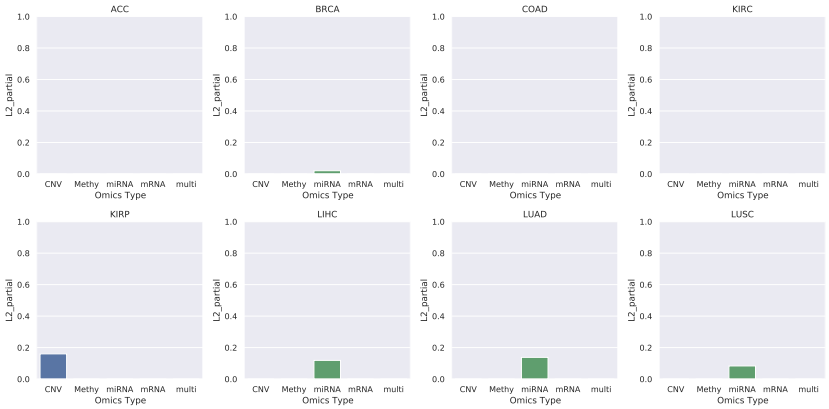
Fonte: O Autor

Figura A.19 – L1_Partial - Classe Yes - Ômicas por tipo de Câncer



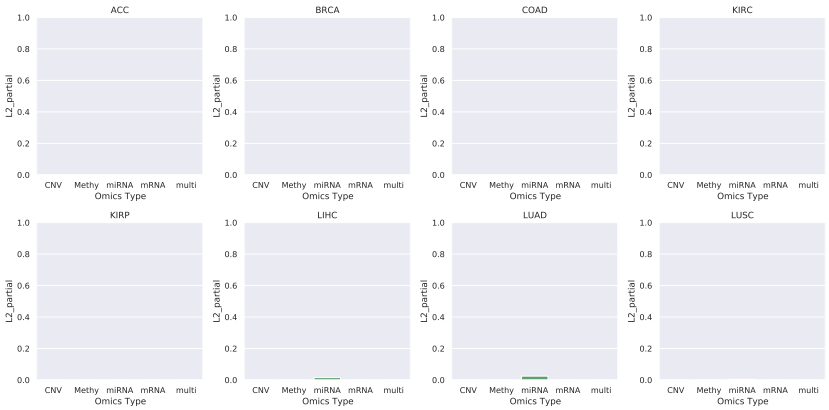
Fonte: O Autor

Figura A.20 – L2_Partial - Classe No - Ômicas por tipo de Câncer



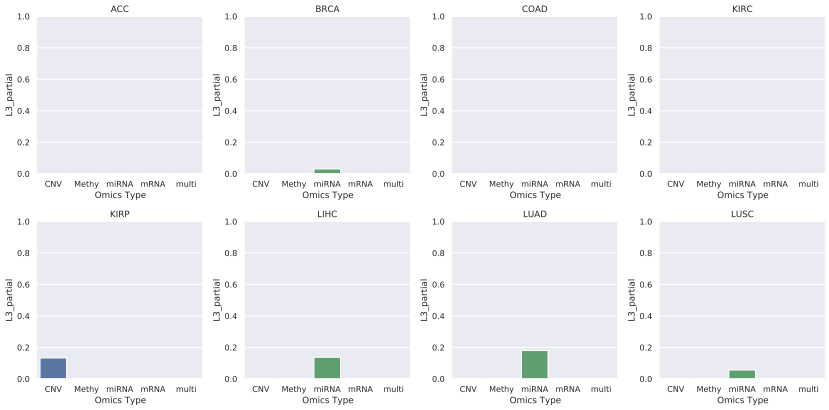
Fonte: O Autor

Figura A.21 – L2_Partial - Classe Yes - Ômicas por tipo de Câncer



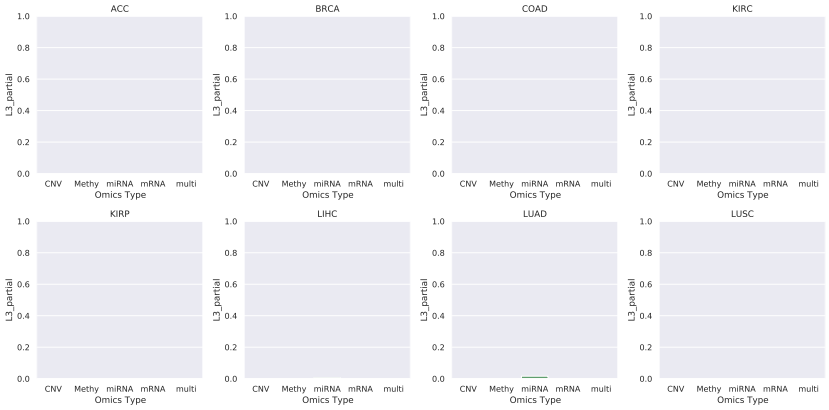
Fonte: O Autor

Figura A.22 – L3_Partial - Classe No - Ômicas por tipo de Câncer



Fonte: O Autor

Figura A.23 – L3_Partial - Classe Yes - Ômicas por tipo de Câncer



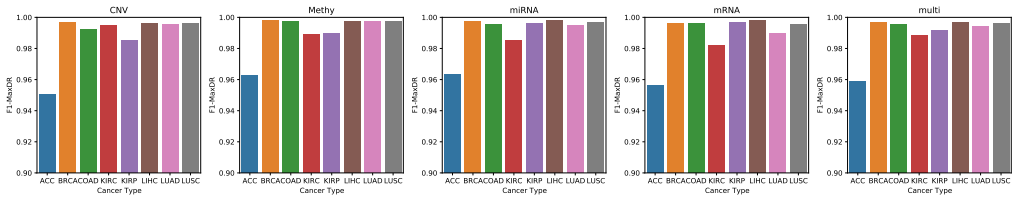
Fonte: O Autor

A.2 Comparação entre Tipos de Câncer por Tipo de Ômica

A.2.1 Medidas de Sobreposição de Atributos

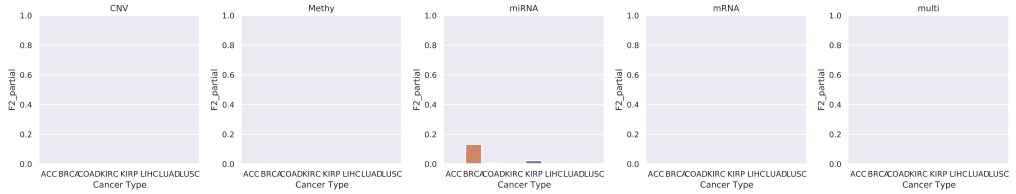
As Figuras A.24 a A.30 representam os resultados comparados dos tipos de câncer para as métricas de sobreposição de atributos.

Figura A.24 – F1-MaxDR - Câncer por tipo de Ômica



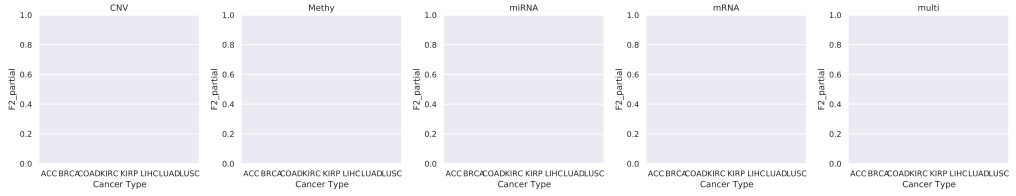
Fonte: O Autor

Figura A.25 – F2_Partial - Classe No - Câncer por tipo de Ômica



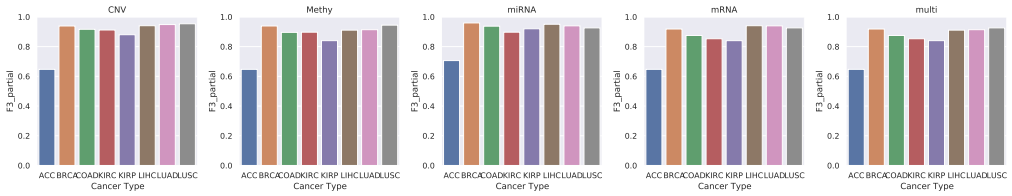
Fonte: O Autor

Figura A.26 – F2_Partial - Classe Yes - Câncer por tipo de Ômica



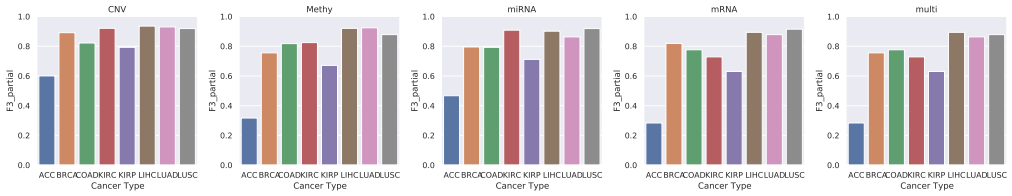
Fonte: O Autor

Figura A.27 – F3_Partial - Classe No - Câncer por tipo de Ômica



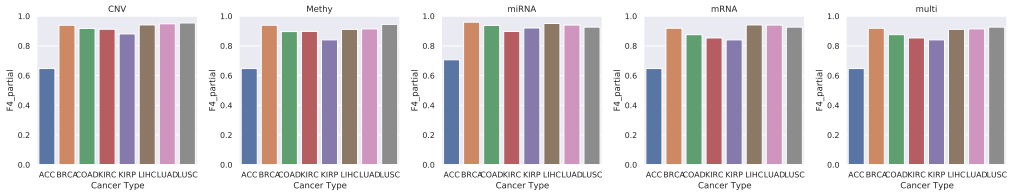
Fonte: O Autor

Figura A.28 – F3_Partial - Classe Yes - Câncer por tipo de Ômica



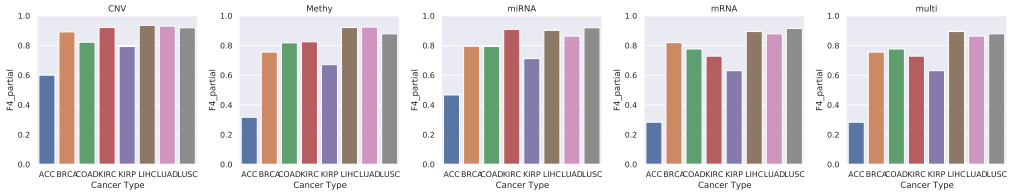
Fonte: O Autor

Figura A.29 – F4_Partial - Classe No - Câncer por tipo de Ômica



Fonte: O Autor

Figura A.30 – F4_Partial - Classe Yes - Câncer por tipo de Ômica

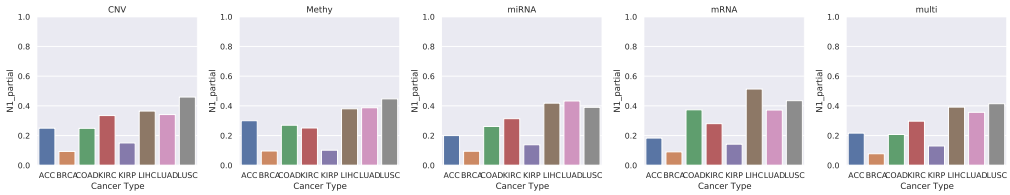


Fonte: O Autor

A.2.2 Medidas de Vizinhança

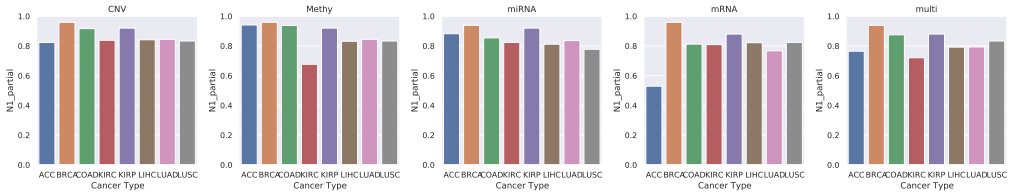
Os resultados das medidas de vizinhança são apresentados das Figuras A.31 a A.40.

Figura A.31 – N1_Partial - Classe No - Câncer por tipo de Ômica



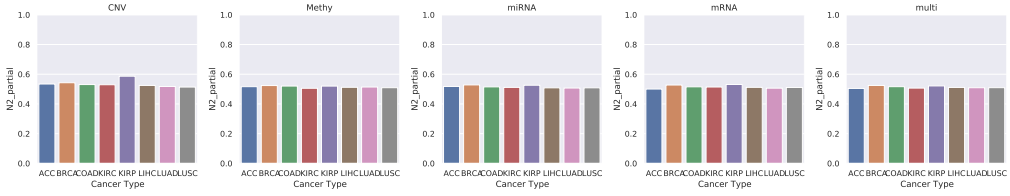
Fonte: O Autor

Figura A.32 – N1_Partial - Classe Yes - Câncer por tipo de Ômica



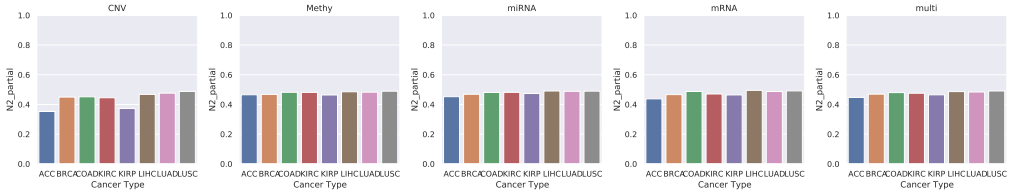
Fonte: O Autor

Figura A.33 – N2_Partial - Classe No - Câncer por tipo de Ômica



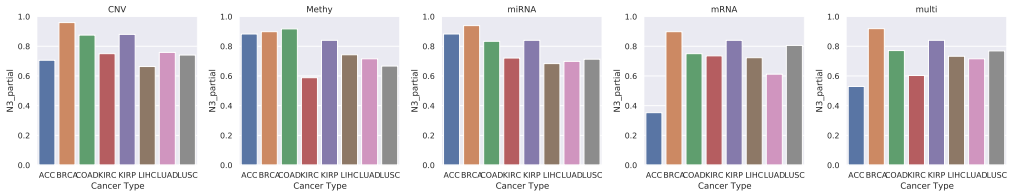
Fonte: O Autor

Figura A.34 – N2_Partial - Classe Yes - Câncer por tipo de Ômica



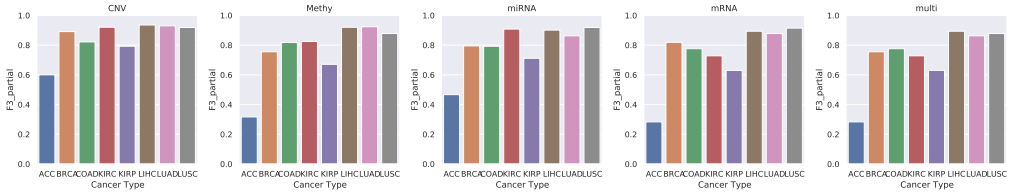
Fonte: O Autor

Figura A.35 – N3_Partial - Classe No - Câncer por tipo de Ômica



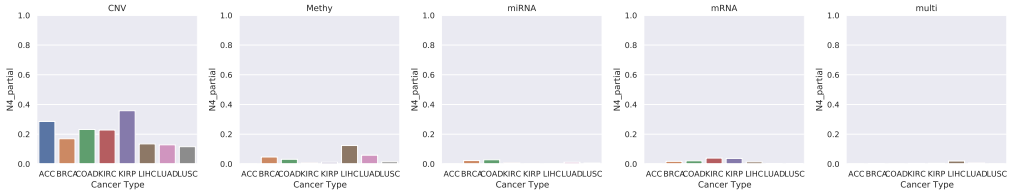
Fonte: O Autor

Figura A.36 – N3_Partial - Classe Yes - Câncer por tipo de Ômica



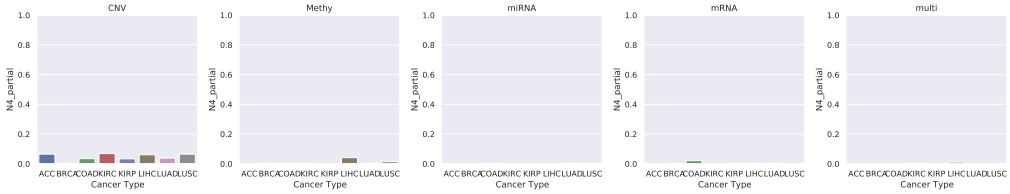
Fonte: O Autor

Figura A.37 – N4_Partial - Classe No - Câncer por tipo de Ômica



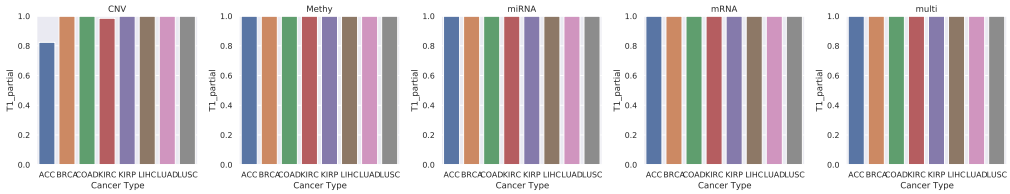
Fonte: O Autor

Figura A.38 – N4_Partial - Classe Yes - Câncer por tipo de Ômica



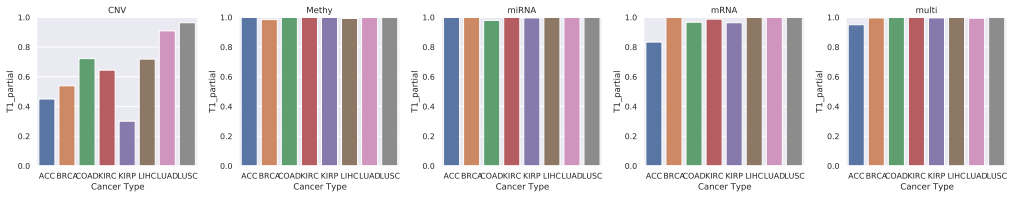
Fonte: O Autor

Figura A.39 – T1_Partial - Classe No - Câncer por tipo de Ômica



Fonte: O Autor

Figura A.40 – T1_Partial - Classe Yes - Câncer por tipo de Ômica

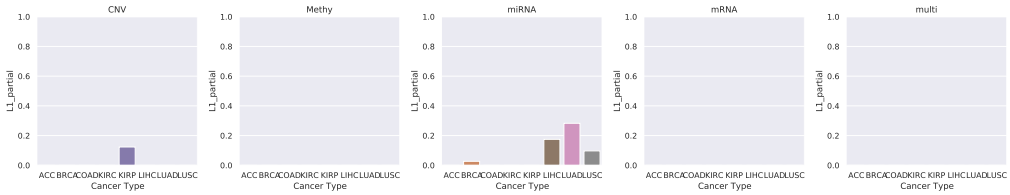


Fonte: O Autor

A.2.3 Medidas de Linearidade

As Figuras A.41 a A.46 comparam os cânceres em termos das medidas de linearidade.

Figura A.41 – L1_Partial - Classe No - Câncer por tipo de Ômica

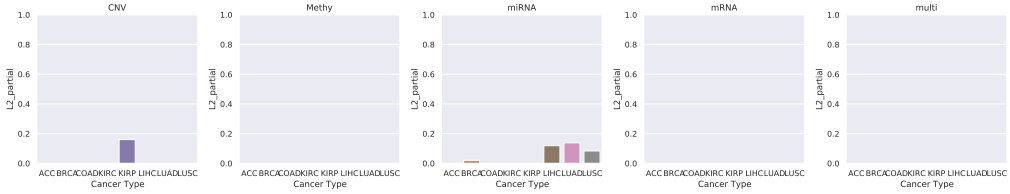


Fonte: O Autor

Figura A.42 – L1_Partial - Classe Yes - Câncer por tipo de Ômica

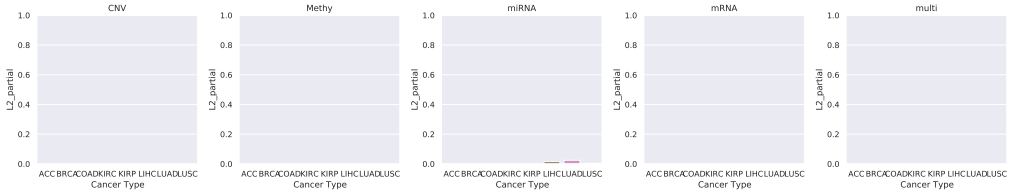
Fonte: O Autor

Figura A.43 – L2_Partial - Classe No - Câncer por tipo de Ômica



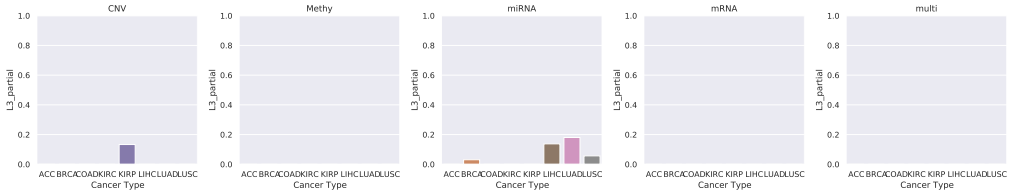
Fonte: O Autor

Figura A.44 – L2_Partial - Classe Yes - Câncer por tipo de Ômica



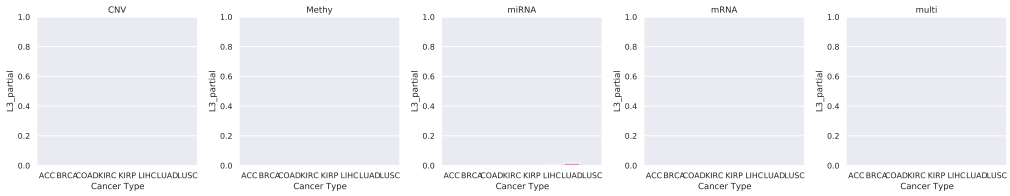
Fonte: O Autor

Figura A.45 – L3_Partial - Classe No - Câncer por tipo de Ômica



Fonte: O Autor

Figura A.46 – L3_Partial - Classe Yes - Câncer por tipo de Ômica



Fonte: O Autor

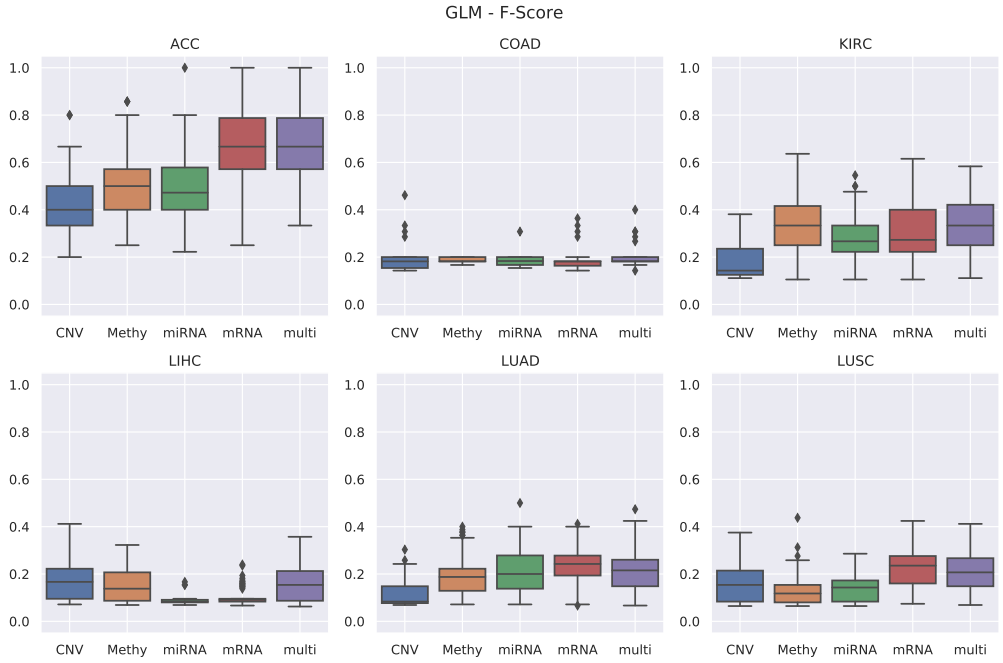
A.3 Métricas de Desempenho na Tarefa de Classificação

Abaixo são colocadas as figuras relativas ao desempenho dos classificadores na tarefa de classificação estudada.

A.3.1 Modelos Lineares Generalizados (GLM)

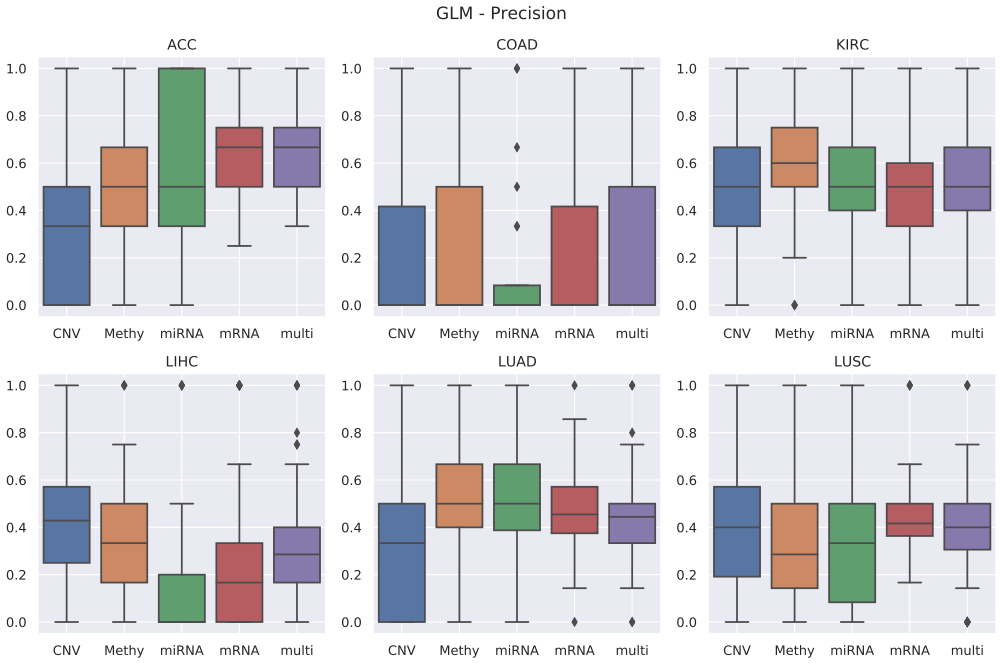
Os resultados para o algoritmo GLM são mostrados nas Figuras A.47 a A.49.

Figura A.47 – F1-Score - Modelo GLM



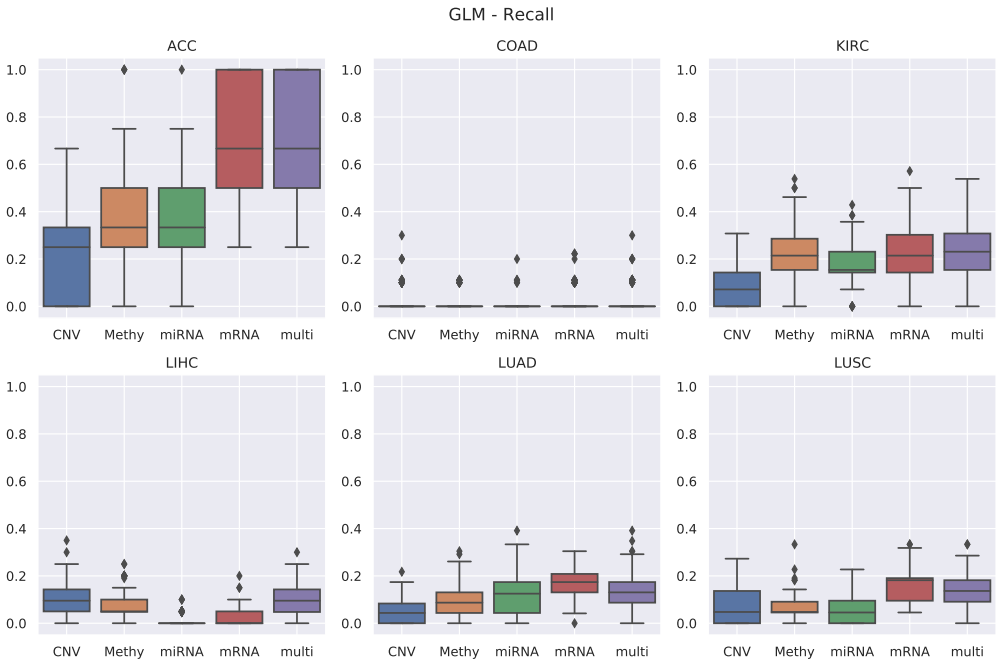
Fonte: O Autor

Figura A.48 – Precisão - Modelo GLM



Fonte: O Autor

Figura A.49 – Recall - Modelo GLM

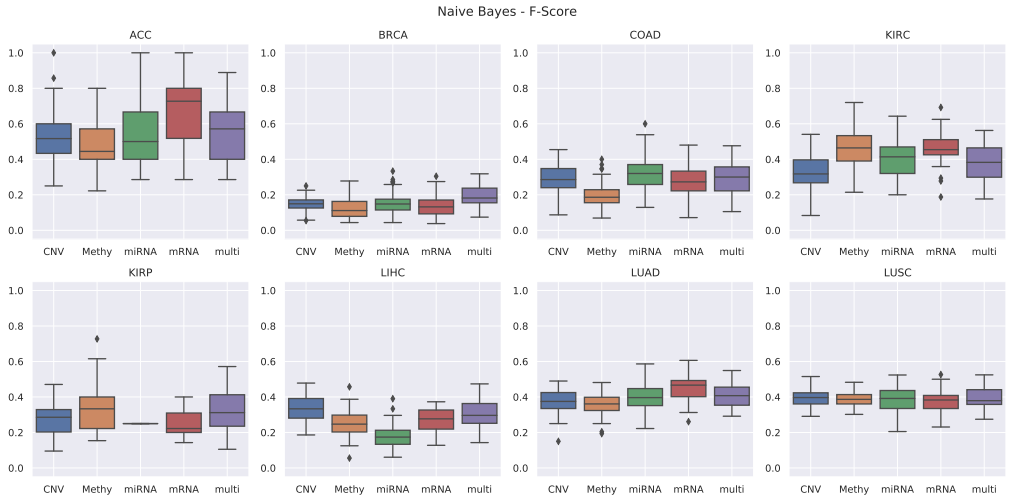


Fonte: O Autor

A.3.2 *Naive-Bayes*

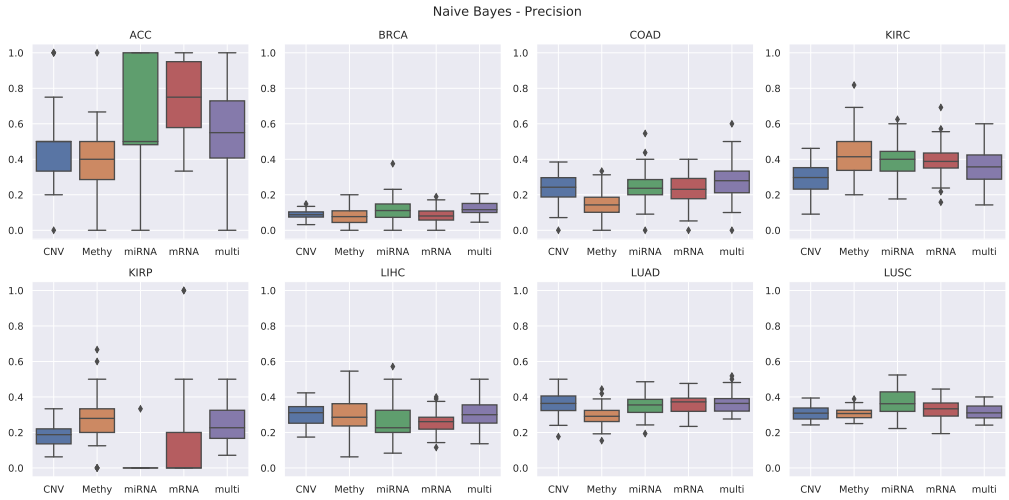
Os resultados para o algoritmo *Naive-Bayes* são mostrados nas Figuras A.50 a A.52.

Figura A.50 – F1-Score - Modelo Naive Bayes



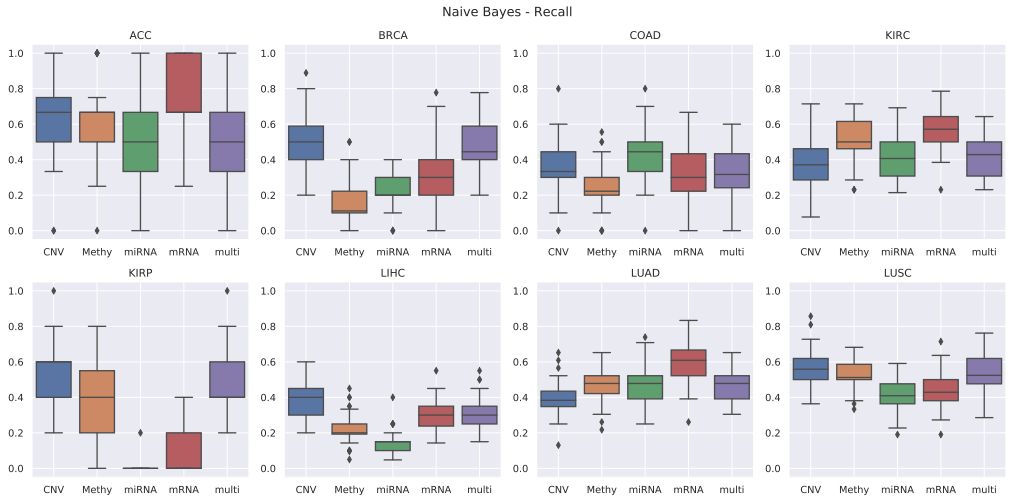
Fonte: O Autor

Figura A.51 – Precisão - Modelo Naive Bayes



Fonte: O Autor

Figura A.52 – Recall - Modelo Naive Bayes



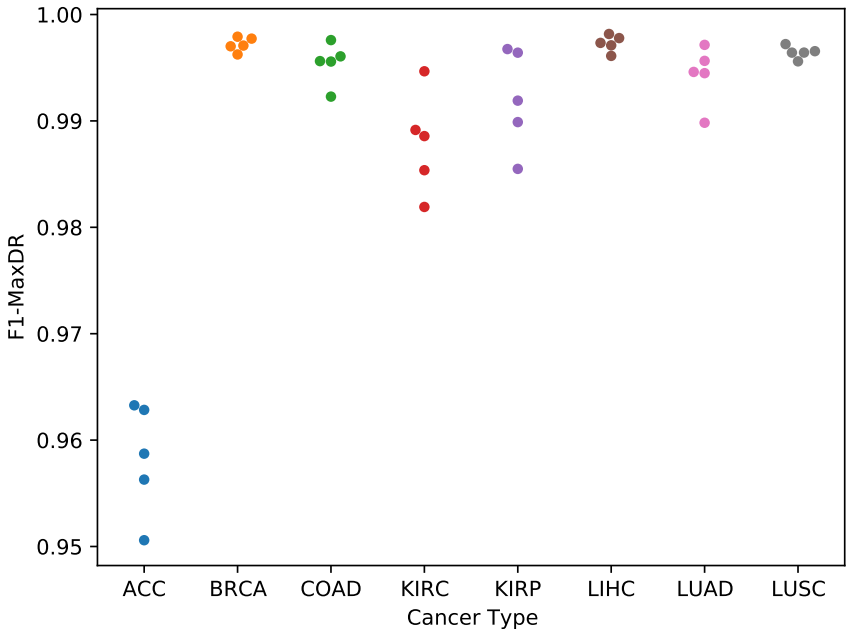
Fonte: O Autor

A.4 Comparação entre as Distribuições de Tipos de Câncer por Métrica de Complexidade

A.4.1 Métricas de Sobreposição de Atributos

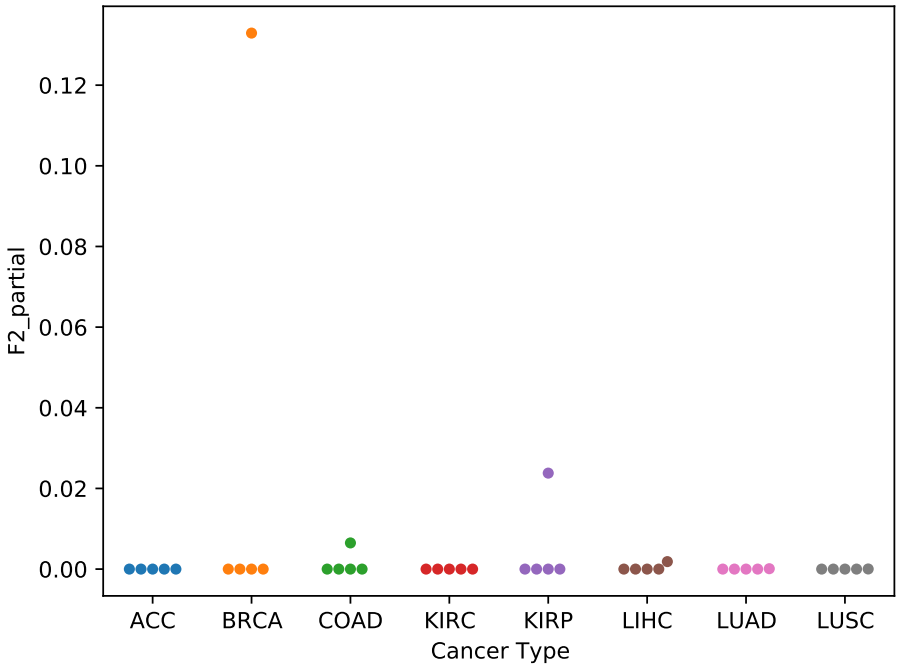
As Figuras A.53 - A.56 mostram *swarmplots* relativos às métricas de sobreposição.

Figura A.53 – F1-MaxDR - Swarmplot - Ômicas por tipo de câncer



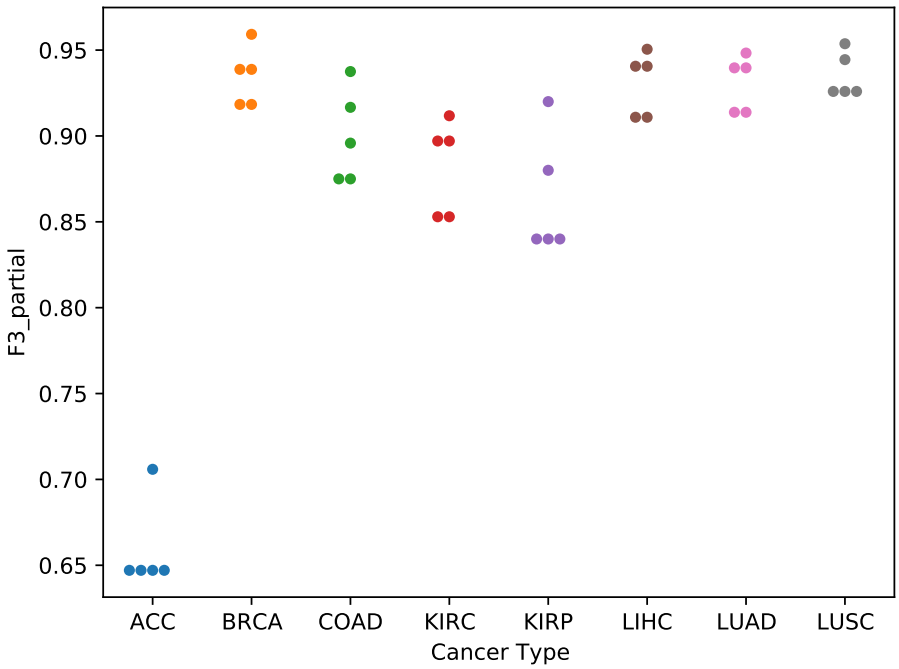
Fonte: O Autor

Figura A.54 – F2_Partial - Swarmplot - Ômicas por tipo de câncer



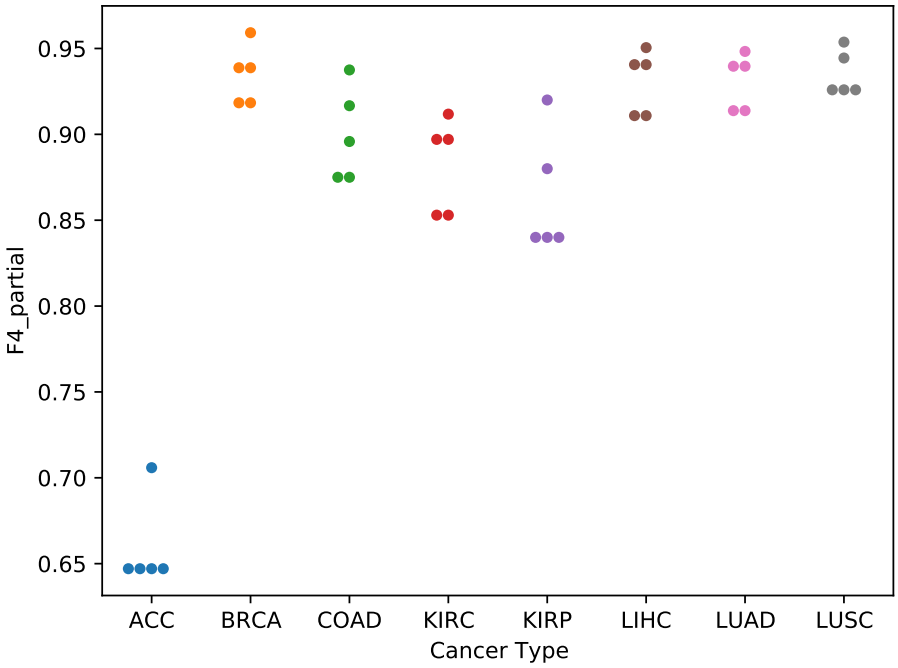
Fonte: O Autor

Figura A.55 – F3_Partial - Swarmplot - Ômicas por tipo de câncer



Fonte: O Autor

Figura A.56 – F4_Partial - Swarmplot - Ômicas por tipo de câncer

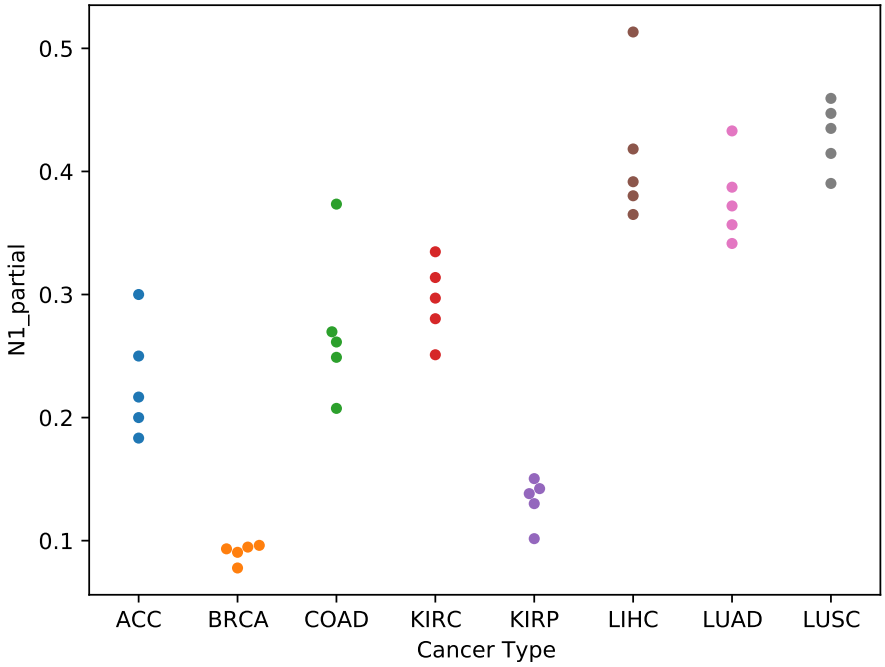


Fonte: O Autor

A.4.2 Métricas de Vizinhança

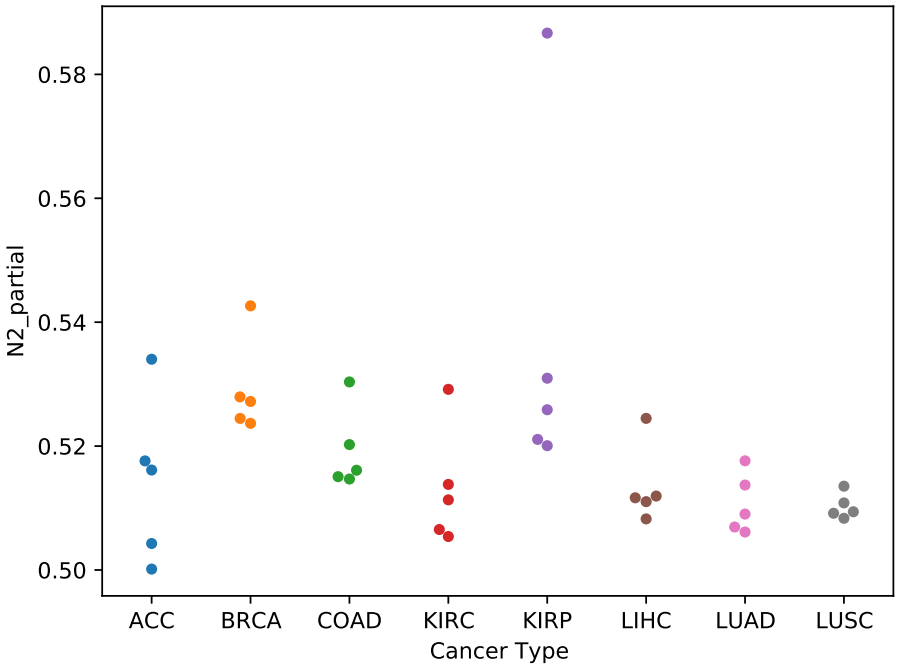
As Figuras de A.57 a A.61 comparam os tipos de câncer em termos das métricas de vizinhança.

Figura A.57 – N1_Partial - Swarmplot - Ômicas por tipo de câncer



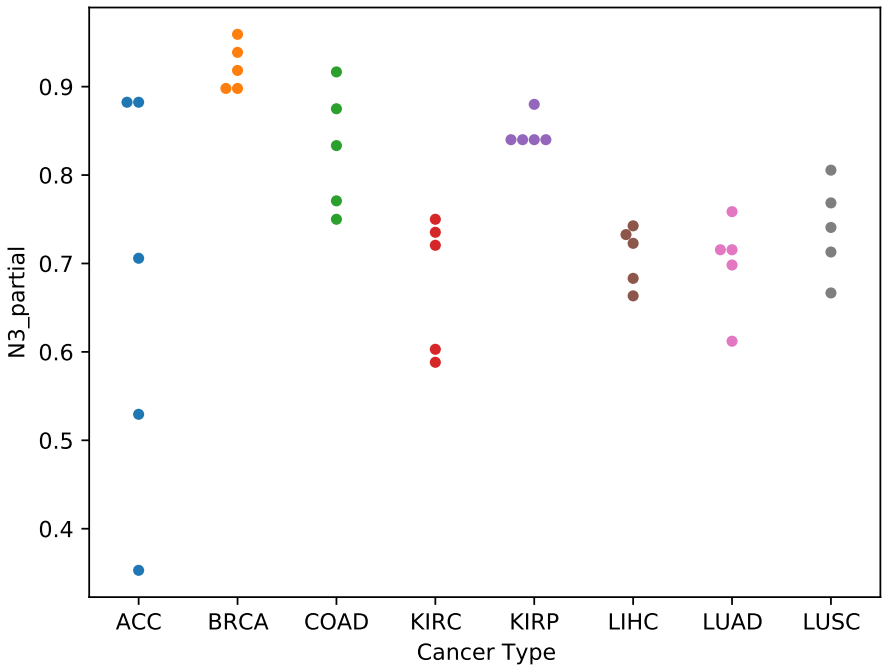
Fonte: O Autor

Figura A.58 – N2_Partial - Swarmplot - Ômicas por tipo de câncer



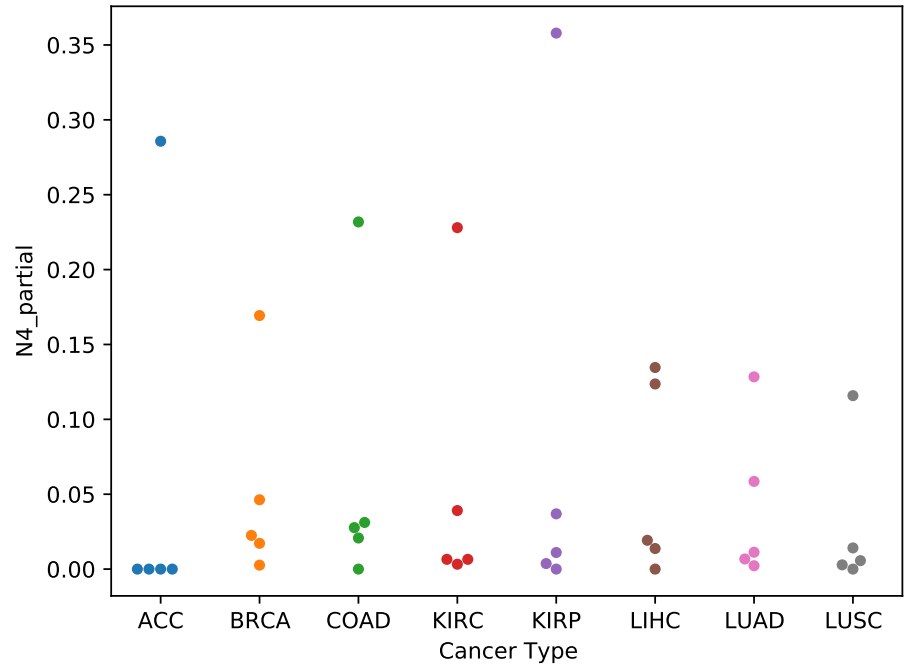
Fonte: O Autor

Figura A.59 – N3_Partial - Swarmplot - Ômicas por tipo de câncer



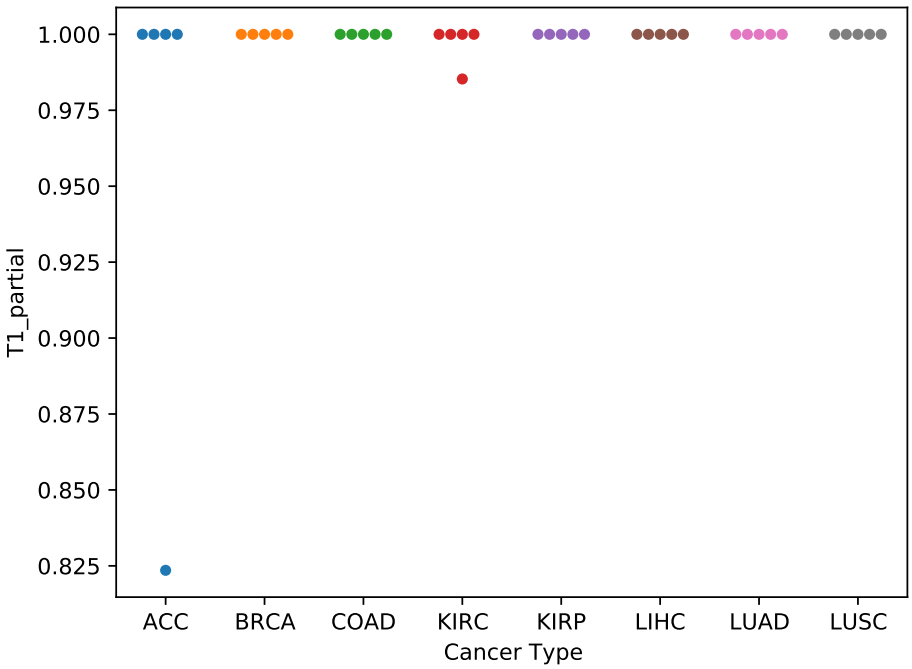
Fonte: O Autor

Figura A.60 – N4_Partial - Swarmplot - Ômicas por tipo de câncer



Fonte: O Autor

Figura A.61 – T1_Partial - Swarmplot - Ômicas por tipo de câncer

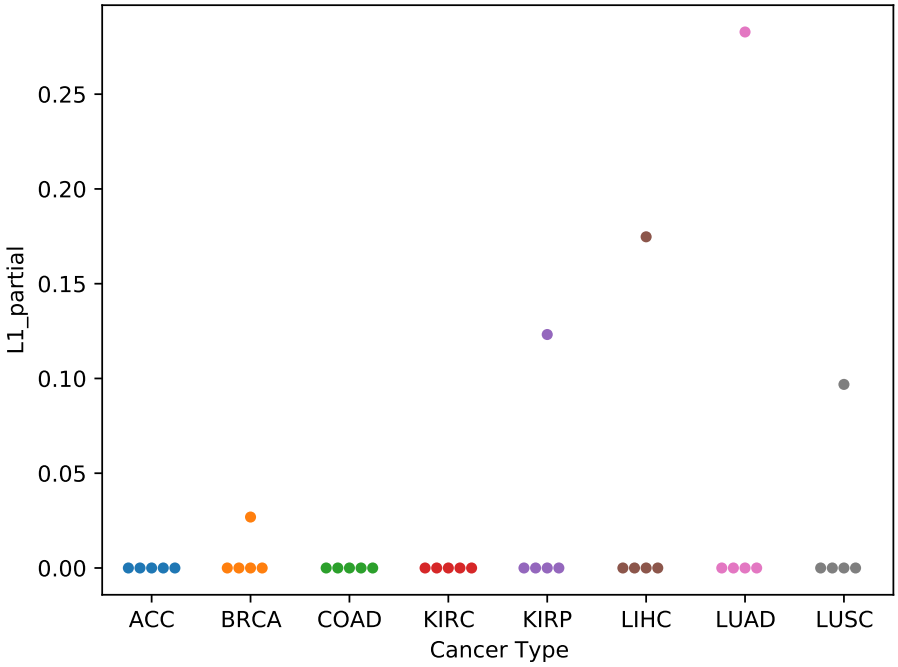


Fonte: O Autor

A.4.3 Métricas de Separabilidade Linear

A comparação entre tipos de câncer é mostrada para as métricas de separabilidade linear da Figura A.62 à A.64.

Figura A.62 – L1_Partial - Swarmplot - Ômicas por tipo de câncer



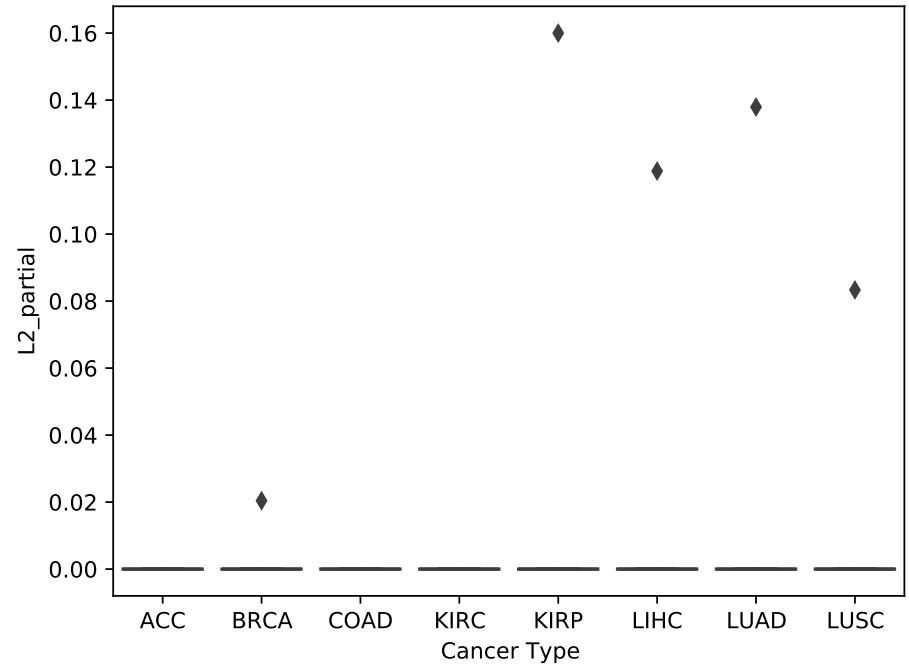
Fonte: O Autor

A.5 Comparação entre as Distribuições de Tipos de Ômica por Métrica de Complexidade

A.5.1 Métricas de Sobreposição

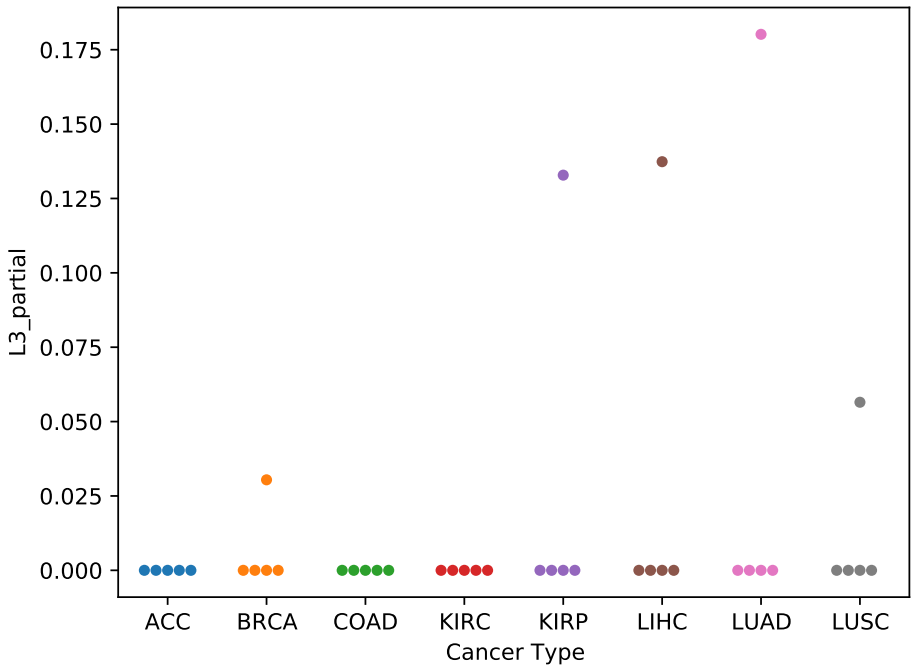
A comparação entre tipos de ômicas por métrica de complexidade é mostrada das Figuras **??** a A.68.

Figura A.63 – L2_Partial - Swarmplot - Ômicas por tipo de câncer



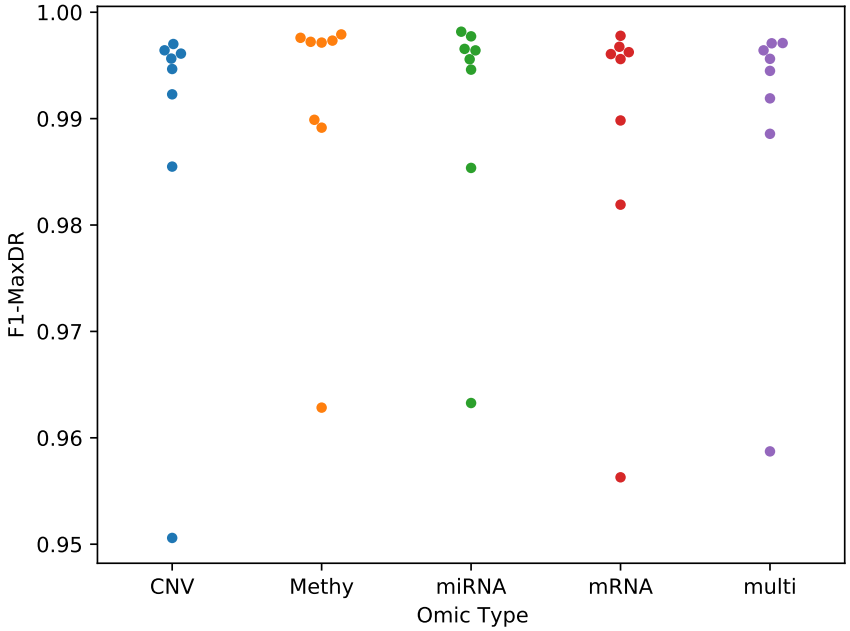
Fonte: O Autor

Figura A.64 – L3_Partial - Swarmplot - Ômicas por tipo de câncer



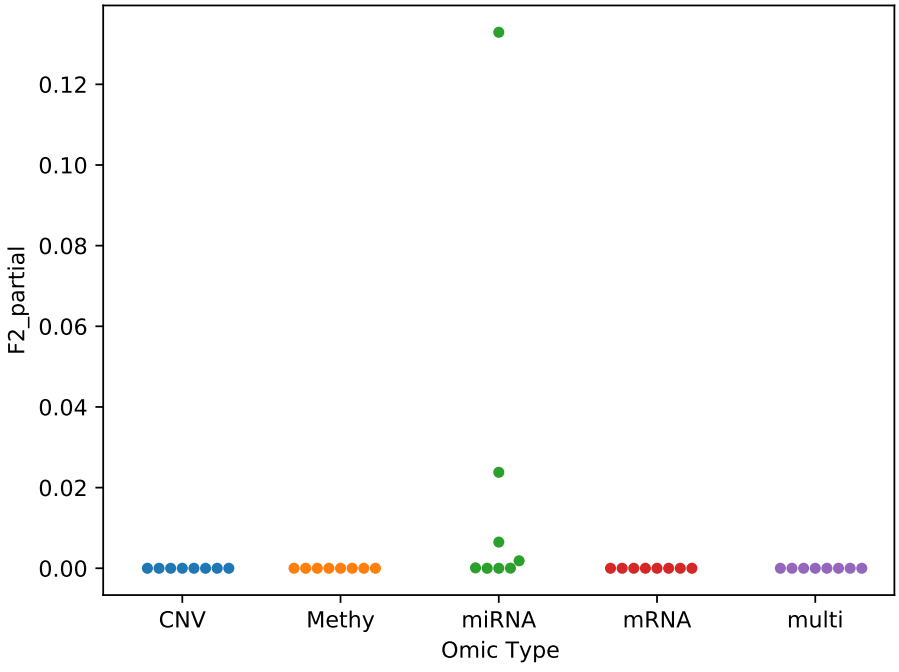
Fonte: O Autor

Figura A.65 – F1-MaxDR - Swarmplot - Tipos de câncer por ômica



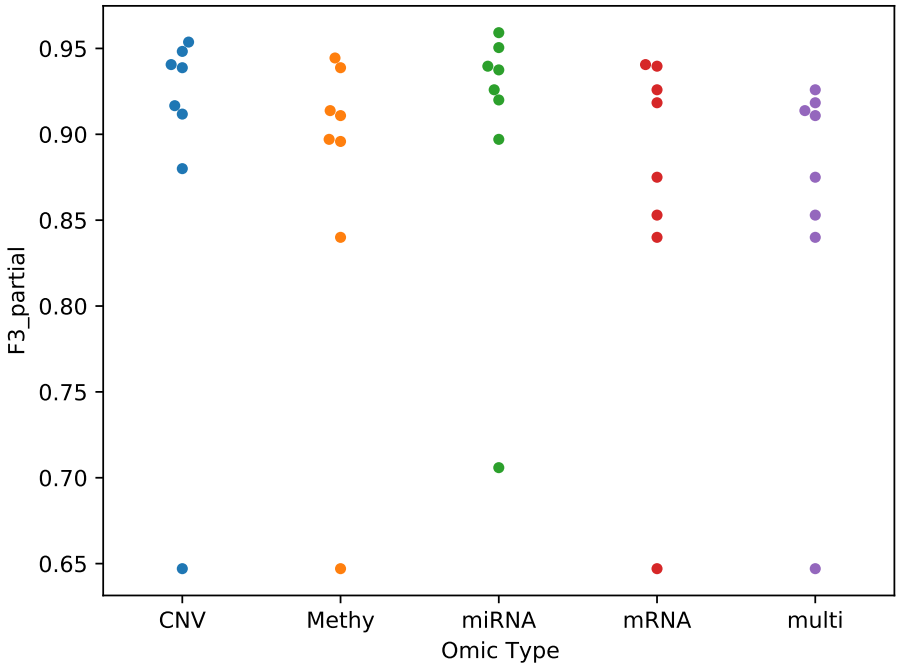
Fonte: O Autor

Figura A.66 – F2_Partial - Swarmplot - Tipos de câncer por ômica



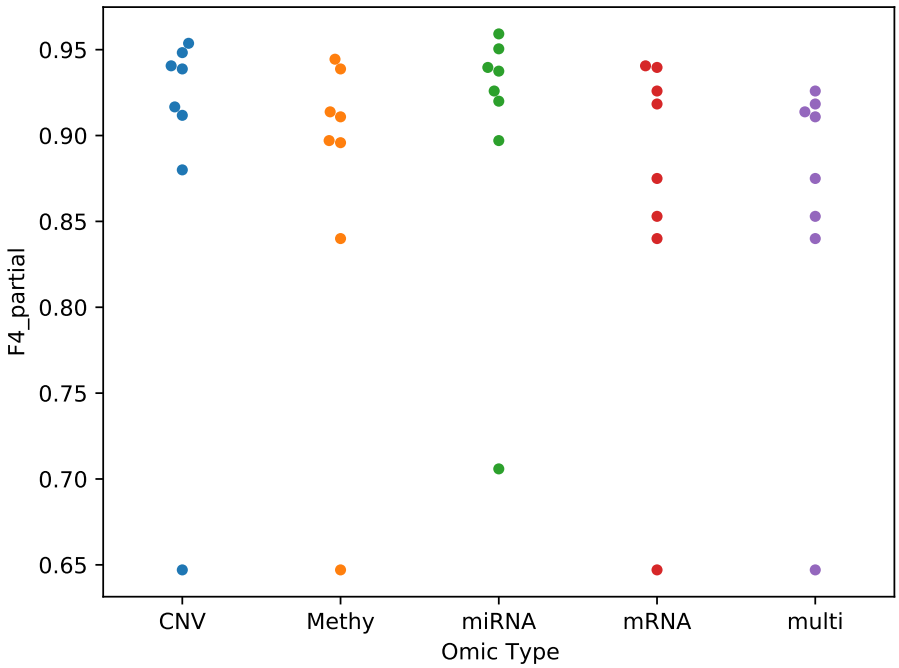
Fonte: O Autor

Figura A.67 – F3_Partial - Swarmplot - Tipos de câncer por ômica



Fonte: O Autor

Figura A.68 – F4_Partial - Swarmplot - Tipos de câncer por ômica

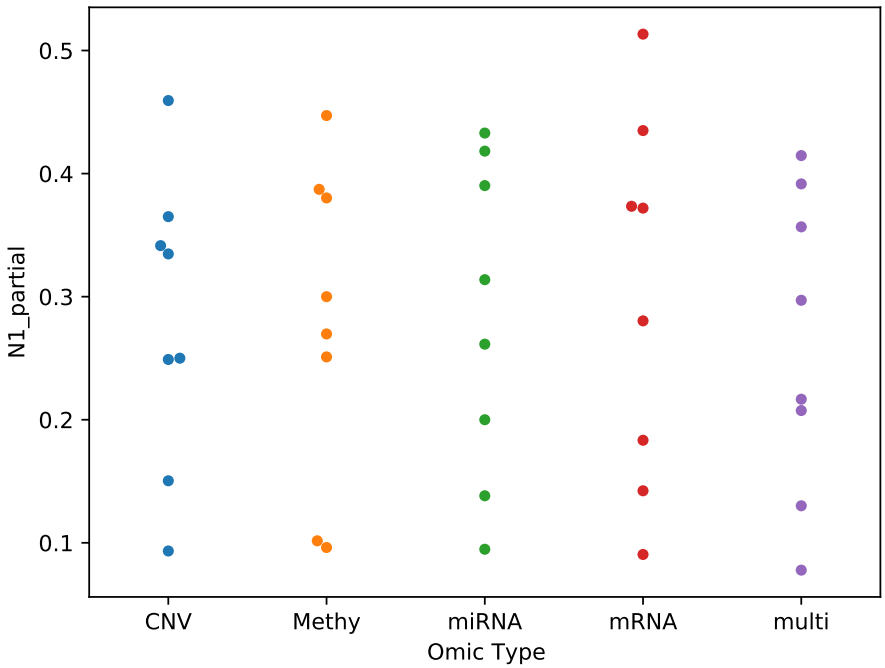


Fonte: O Autor

A.5.2 Métricas de Vizinhança

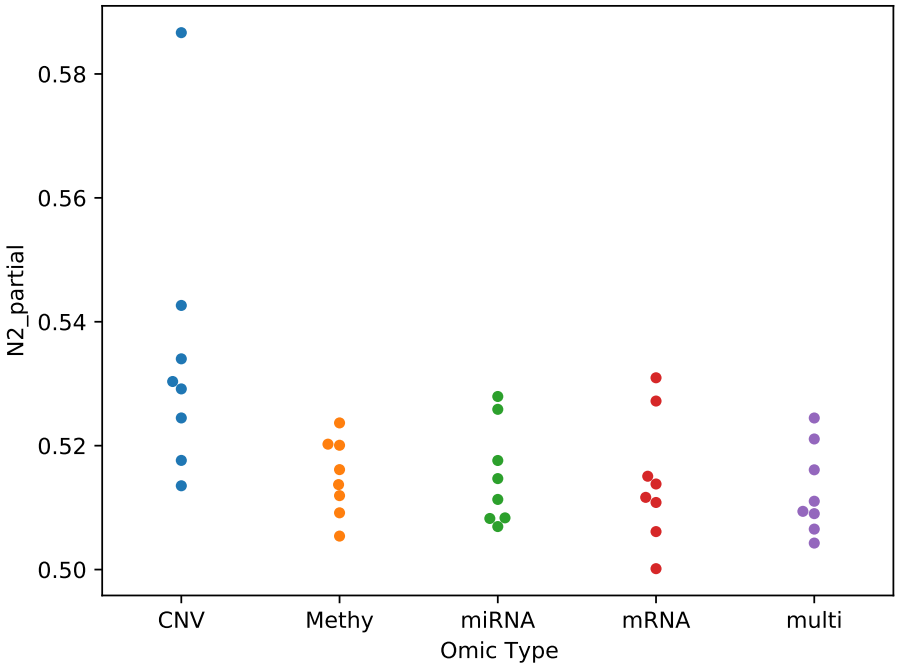
A comparação entre tipos de ômicas por métrica de complexidade é mostrada das Figuras **??** a A.73.

Figura A.69 – N1_Partial - Swarmplot - Tipos de câncer por ômica



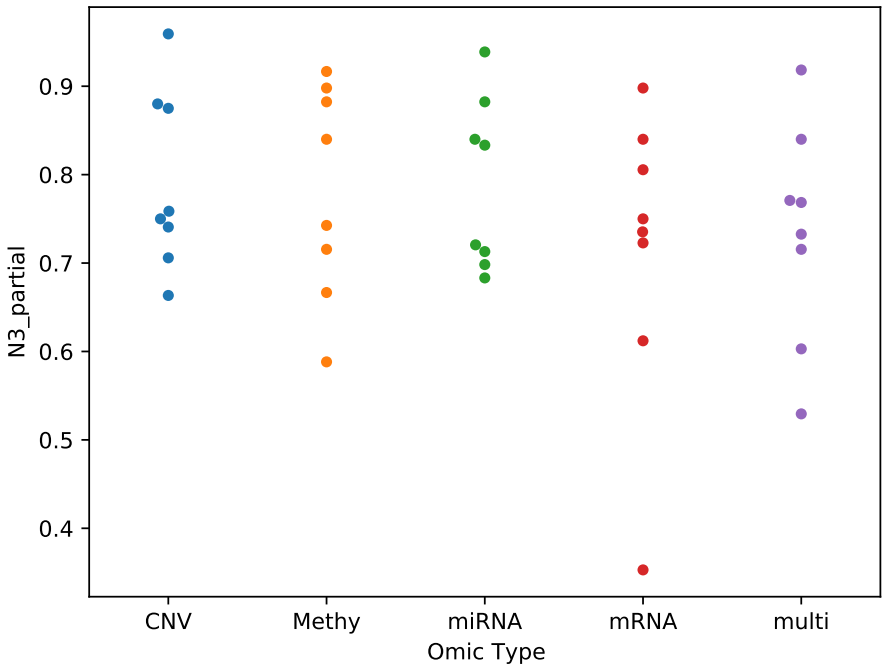
Fonte: O Autor

Figura A.70 – N2_Partial - Swarmplot - Tipos de câncer por ômica



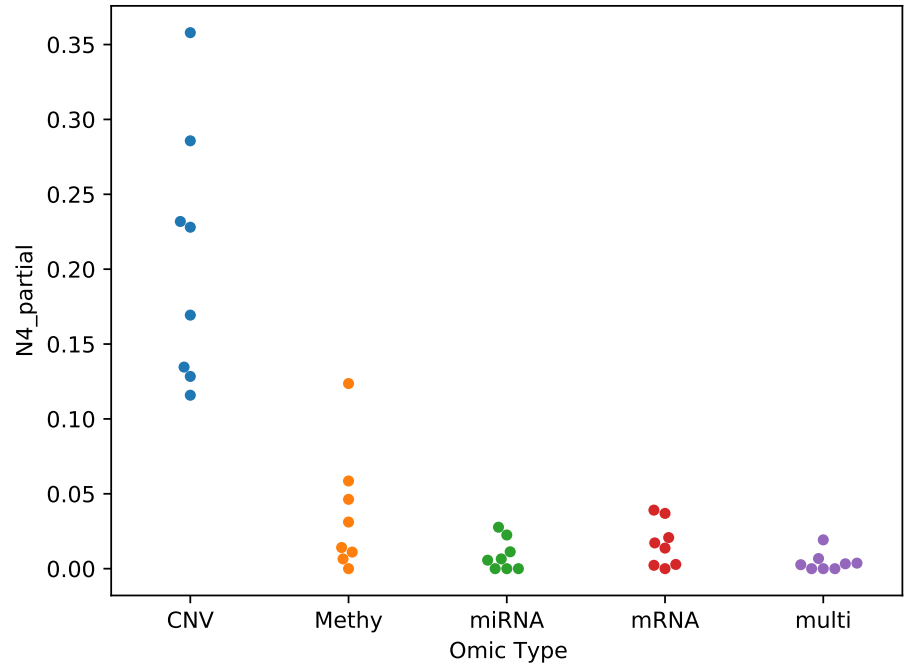
Fonte: O Autor

Figura A.71 – N3_Partial - Swarmplot - Tipos de câncer por ômica



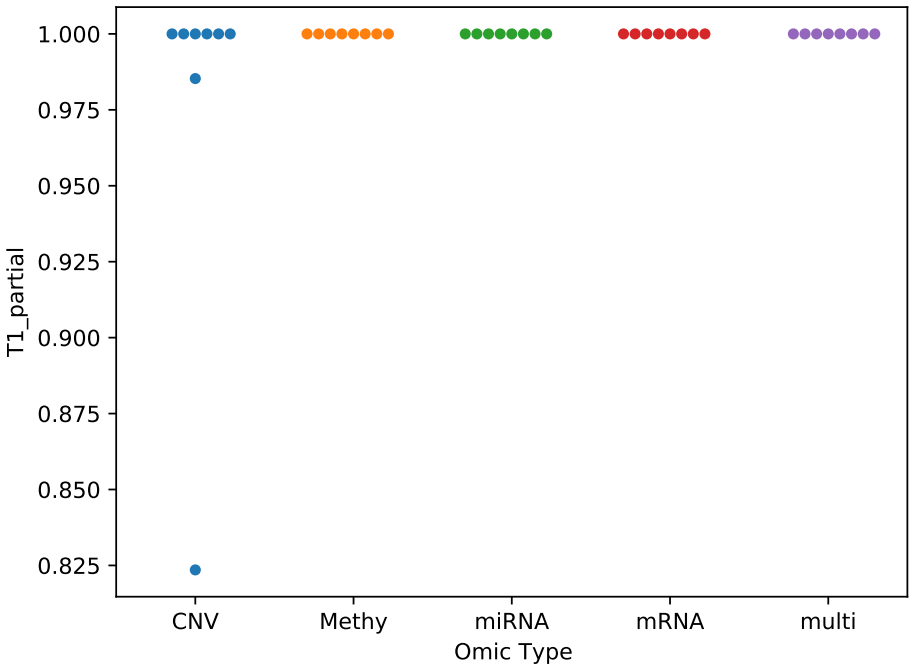
Fonte: O Autor

Figura A.72 – N4_Partial - Swarmplot - Tipos de câncer por ômica



Fonte: O Autor

Figura A.73 – T1_Partial - Swarmplot - Tipos de câncer por ômica

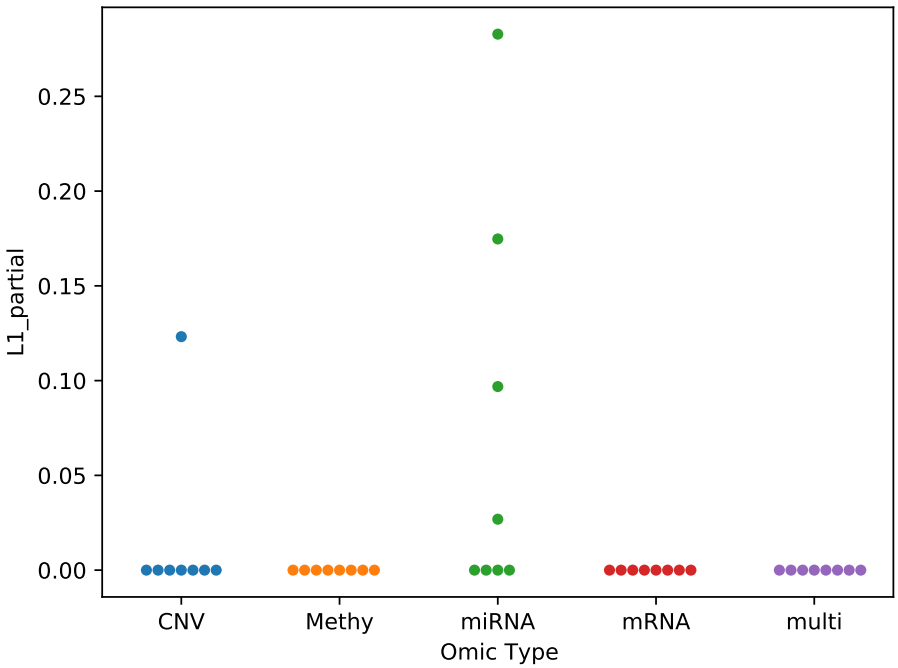


Fonte: O Autor

A.5.3 Métricas de Linearidade

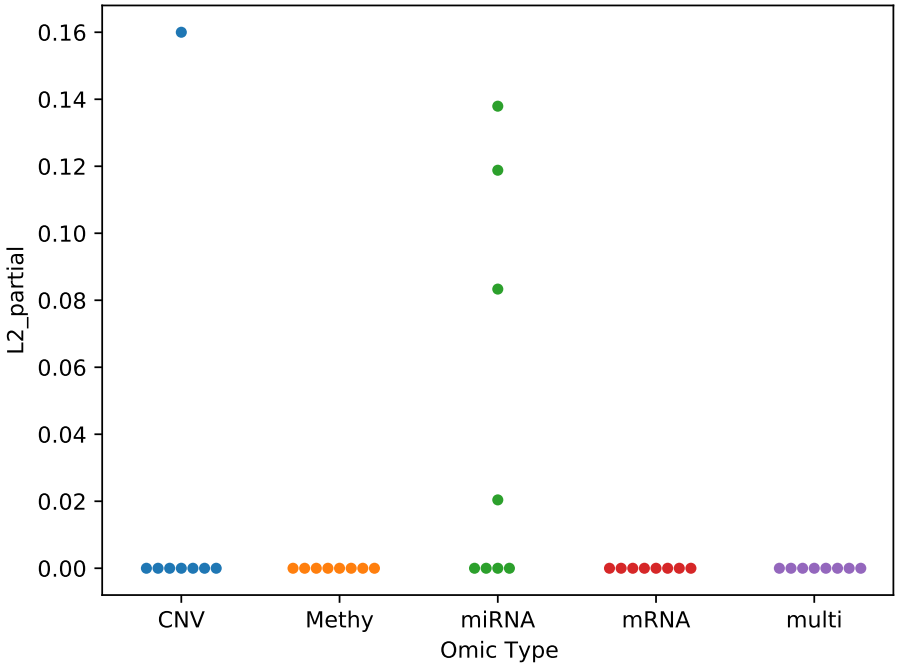
A comparação entre tipos de ômicas por métrica de complexidade é mostrada das Figuras **??** a A.76.

Figura A.74 – L1_Partial - Swarmplot - Tipos de câncer por ômica



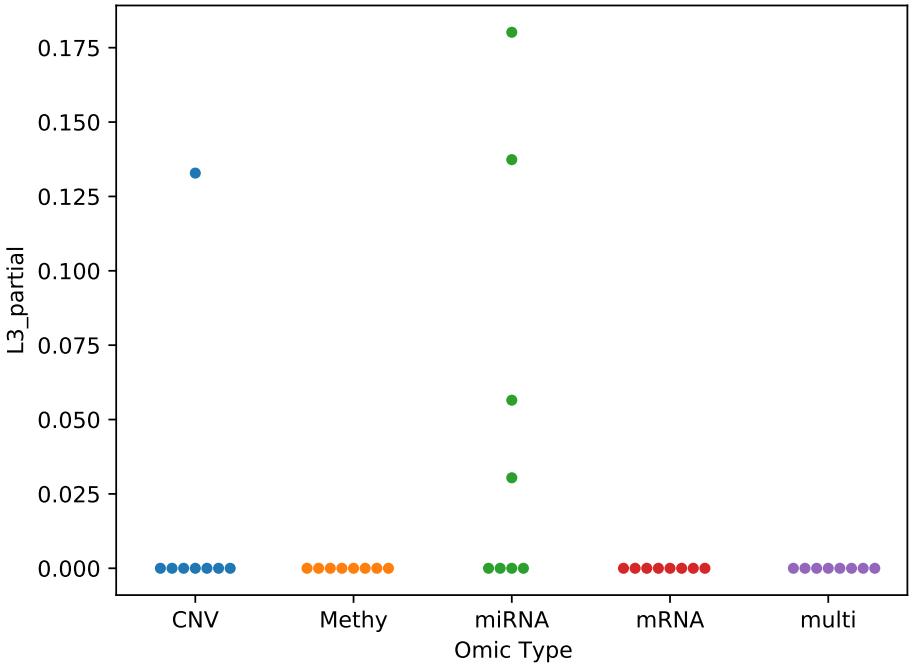
Fonte: O Autor

Figura A.75 – L2_Partial - Swarmplot - Tipos de câncer por ômica



Fonte: O Autor

Figura A.76 – L3_Partial - Swarmplot - Tipos de câncer por ômica



Fonte: O Autor

A.6 Correlações entre Tipos de Câncer

Figura A.77 – Correlação Entre os Tipos de Câncer - Duas Classes - Todas Medidas de Complexidade

Fonte: O Autor

Figura A.78 – Correlação Entre os Tipos de Câncer - Duas Classes - Medidas de Complexidade com Baixa Correlação

Fonte: O Autor

Figura A.79 – Correlação Entre os Tipos de Câncer - Classe No - Todas Medidas de Complexidade

Fonte: O Autor

Figura A.80 – Correlação Entre os Tipos de Câncer - Classe No - Medidas de Complexidade com Baixa Correlação

Fonte: O Autor

Figura A.81 – Correlação Entre os Tipos de Câncer - Classe Yes - Todas Medidas de Complexidade

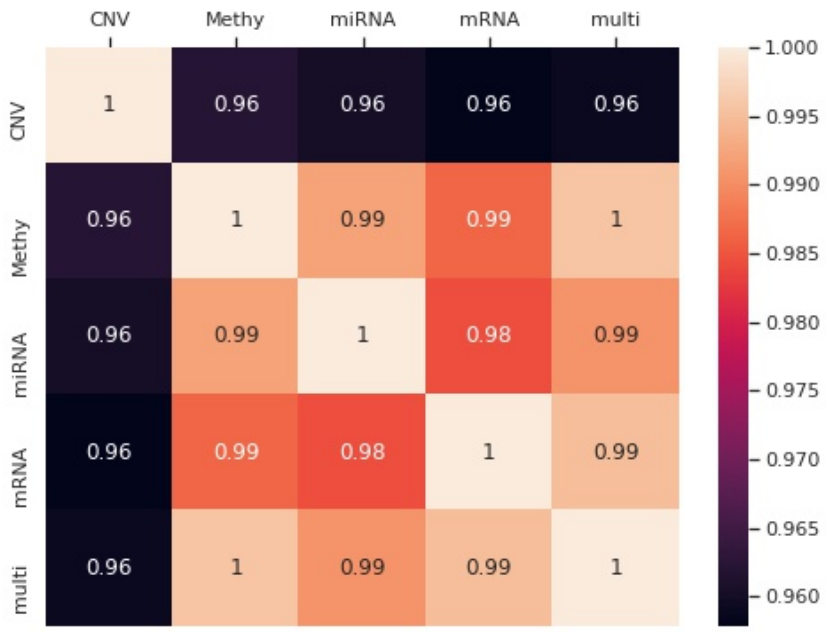
Fonte: O Autor

Figura A.82 – Correlação Entre os Tipos de Câncer - Classe Yes - Medidas de Complexidade com Baixa Correlação

Fonte: O Autor

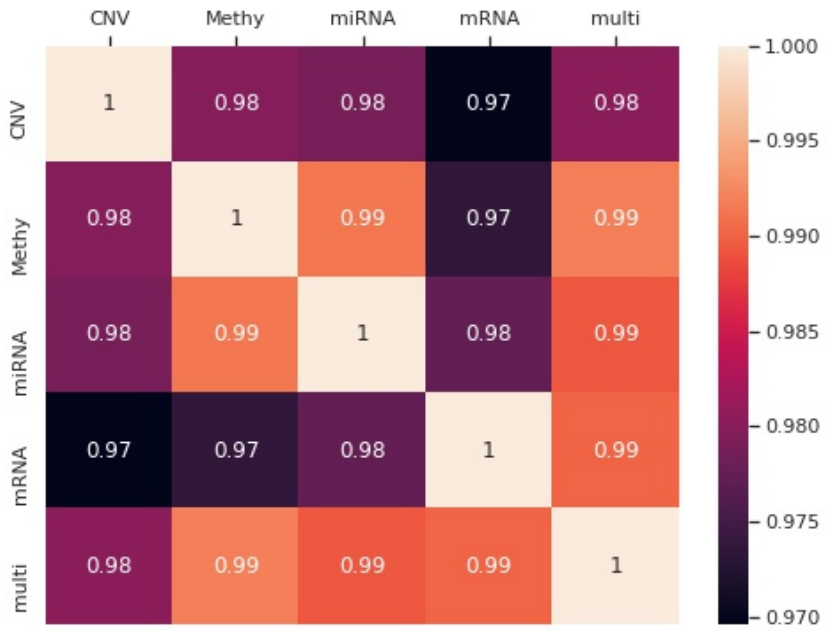
A.7 Correlações entre Tipos de Ômicas

Figura A.83 – Correlação Entre os Tipos de Ômica - Duas Classes - Todas Medidas de Complexidade



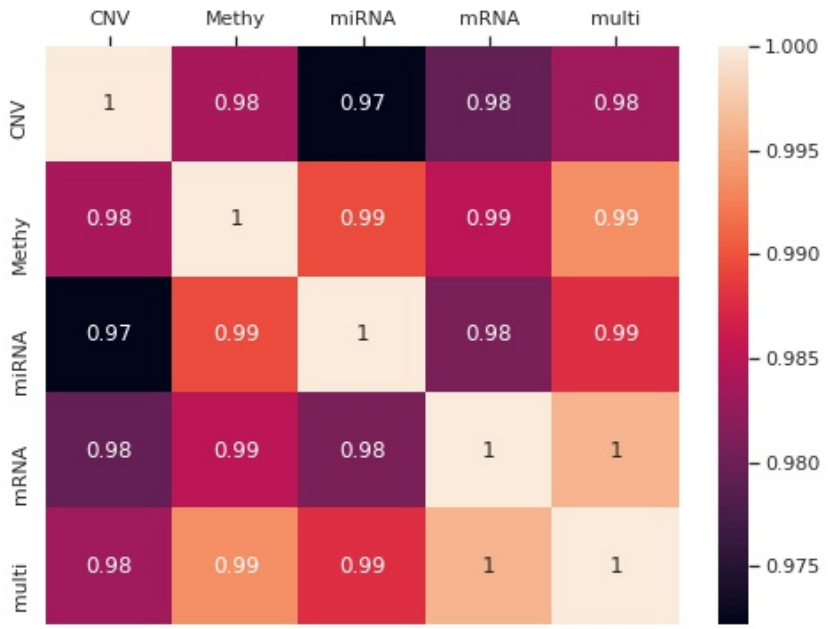
Fonte: O Autor

Figura A.84 – Correlação Entre os Tipos de Ômica - Duas Classes - Medidas de Complexidade com Baixa Correlação



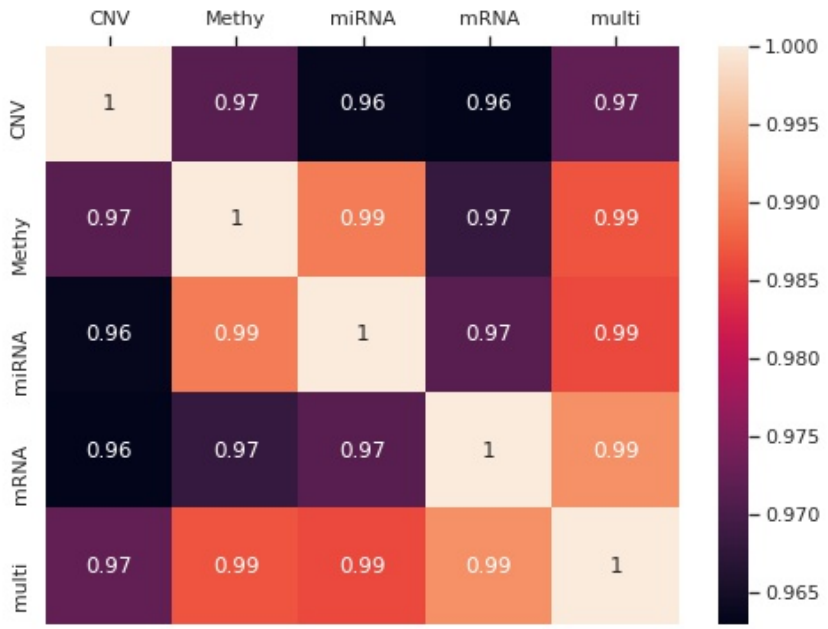
Fonte: O Autor

Figura A.85 – Correlação Entre os Tipos de Ômica - Classe No - Todas Medidas de Complexidade



Fonte: O Autor

Figura A.86 – Correlação Entre os Tipos de Ômica - Classe No - Medidas de Complexidade com Baixa Correlação

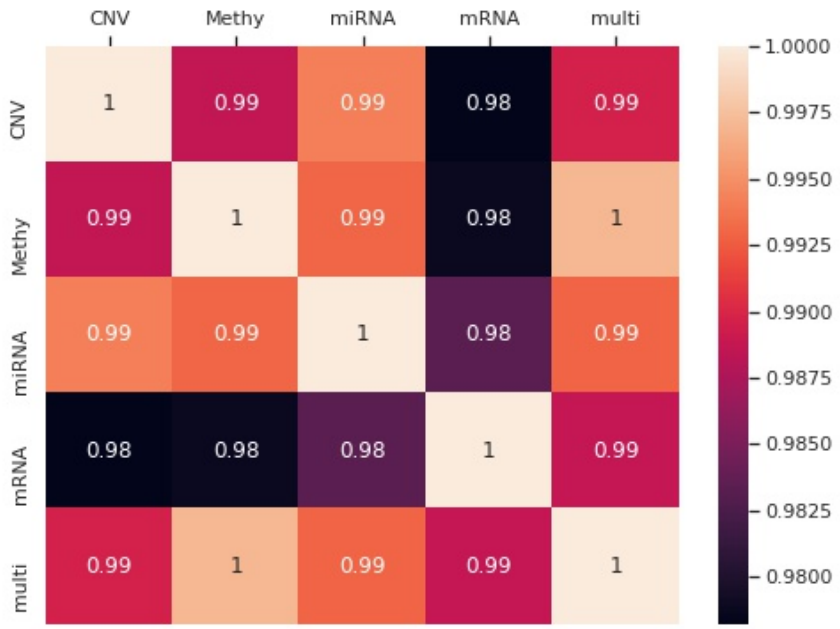


Fonte: O Autor

Figura A.87 – Correlação Entre os Tipos de Ômica - Classe Yes - Todas Medidas de Complexidade

Fonte: O Autor

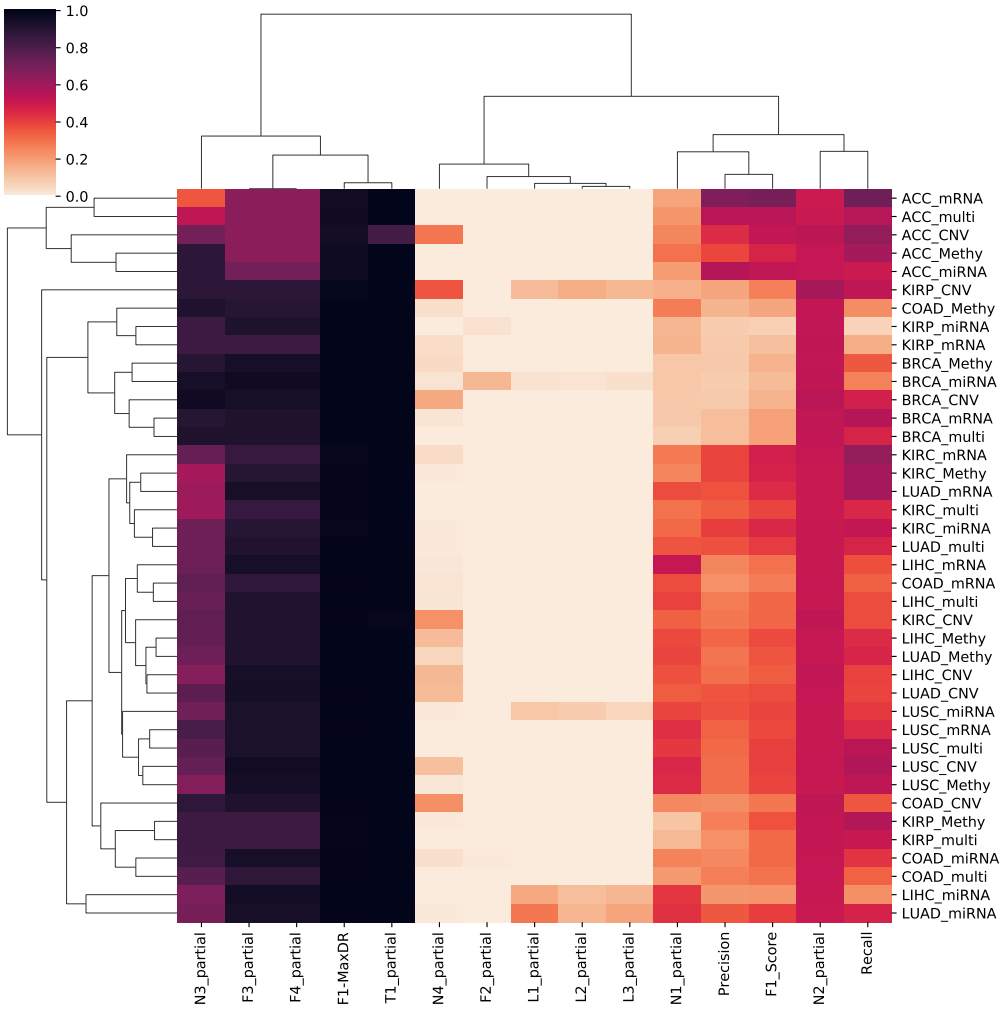
Figura A.88 – Correlação Entre os Tipos de Ômica - Classe Yes - Medidas de Complexidade com Baixa Correlação



Fonte: O Autor

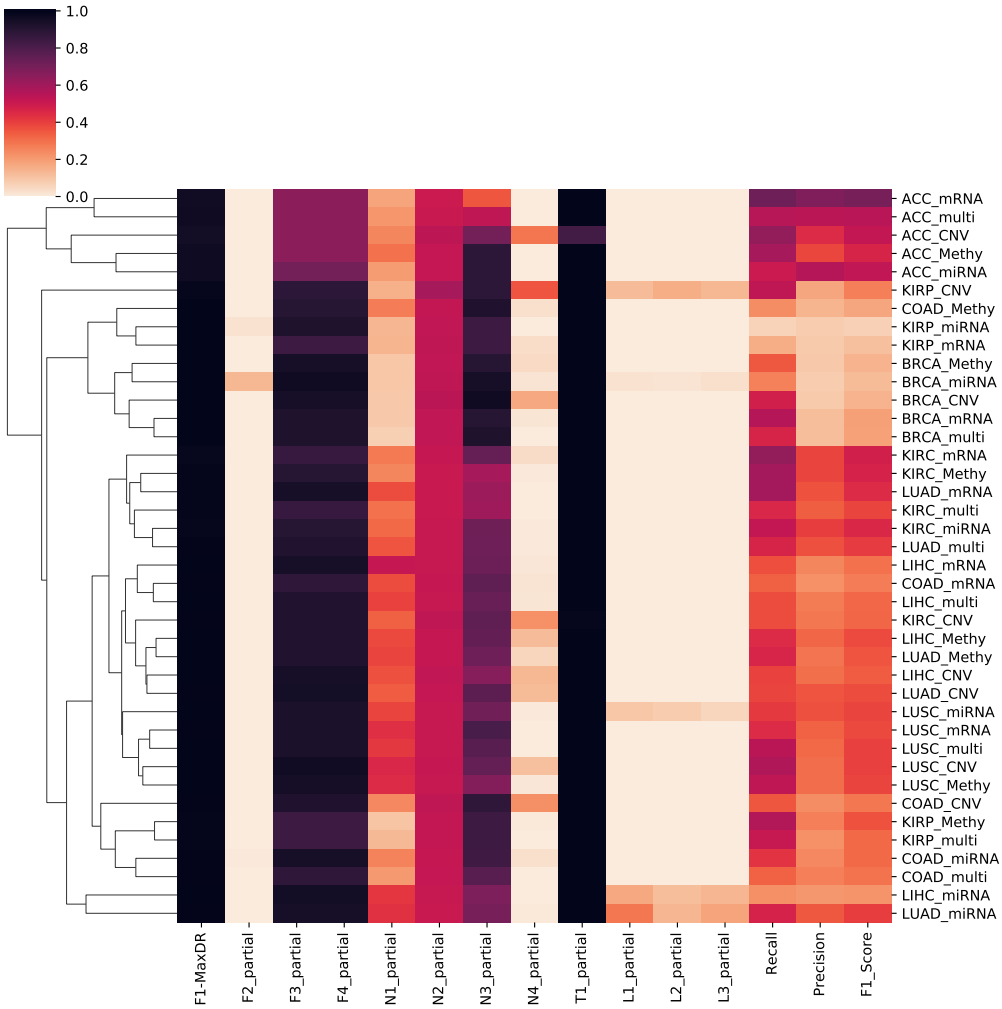
A.8 Mapas de Calor com Agrupamento Hierárquico

Figura A.89 – Clustermap Para a Classe No - Modelo Naive Bayes - Todas as Medidas



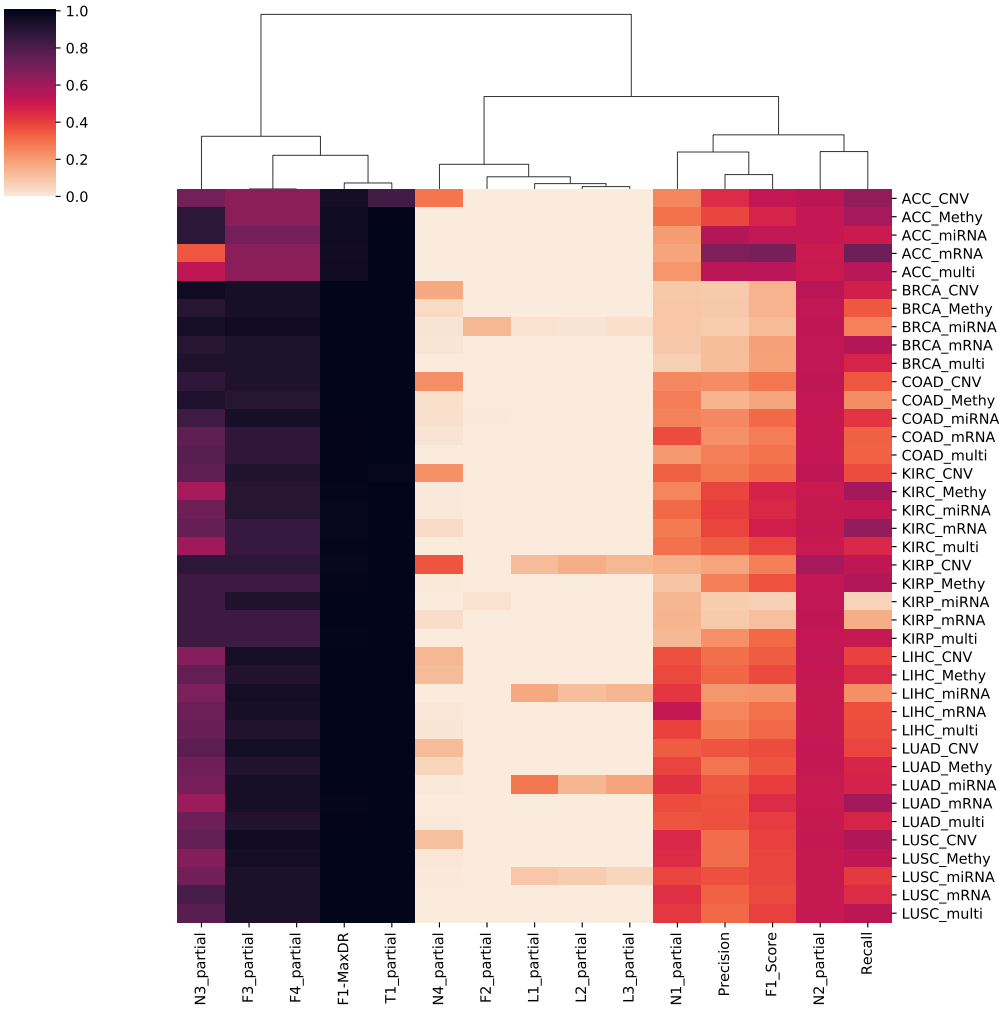
Fonte: O Autor

Figura A.90 – Clustermap Para a Classe No - Modelo Naive Bayes - Por Conjuntos de Dados - Todas Medidas



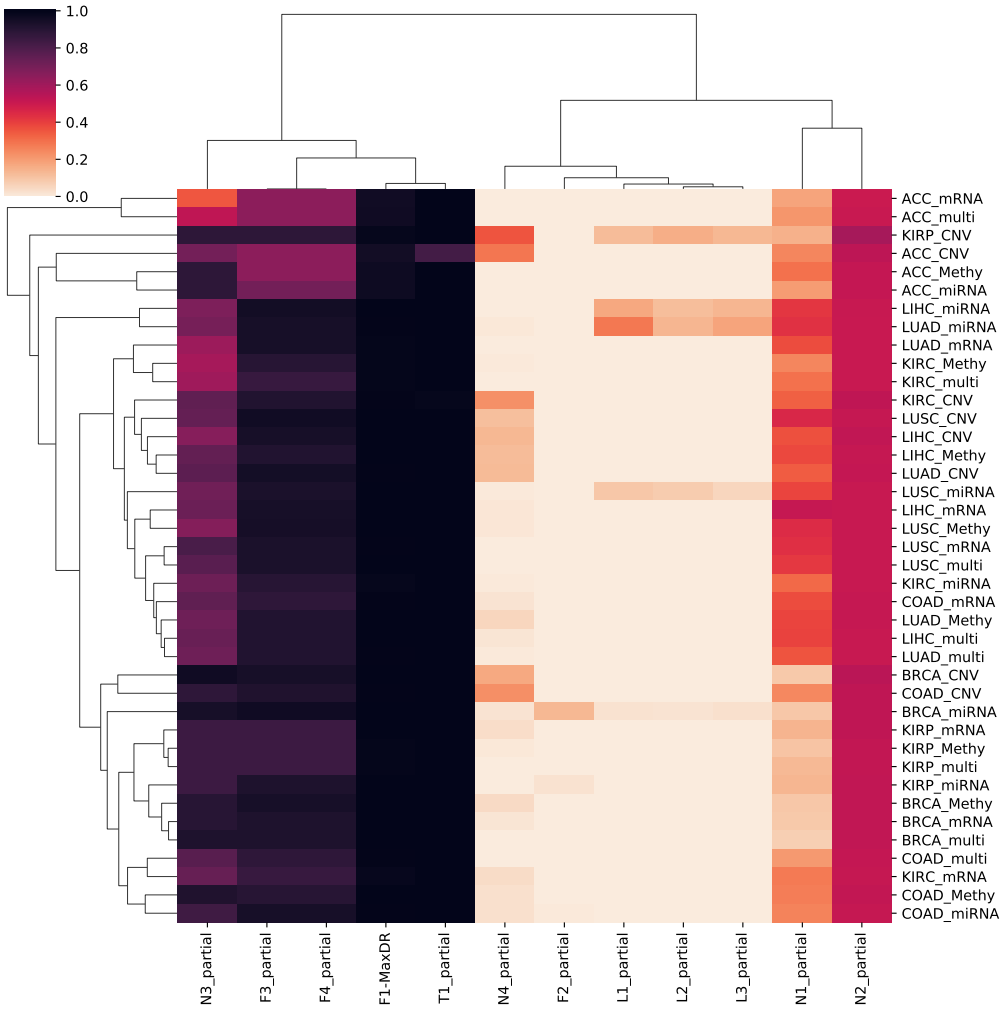
Fonte: O Autor

Figura A.91 – Clustermap Para a Classe No - Modelo Naive Bayes - Por Medida - Todas Medidas



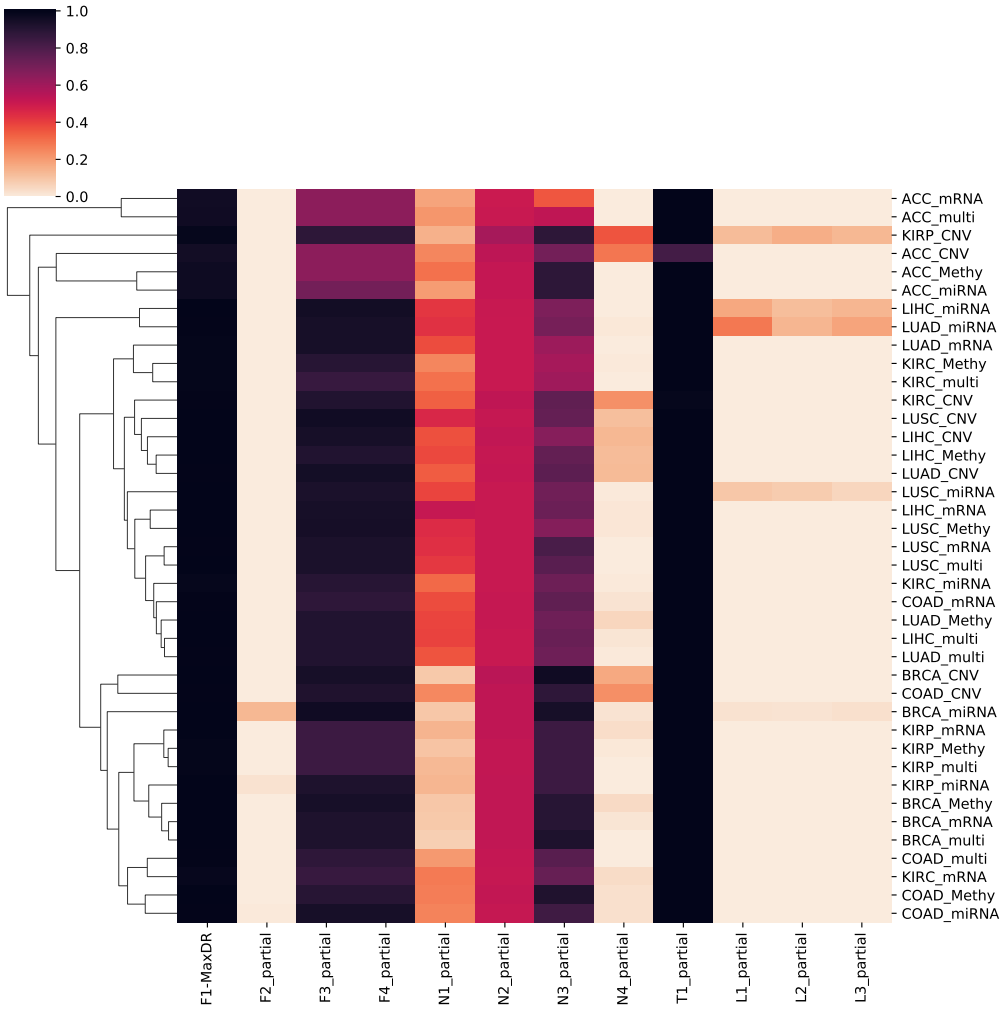
Fonte: O Autor

Figura A.92 – Clustermap Para a Classe No - Modelo Naive Bayes - Complexidade Apenas



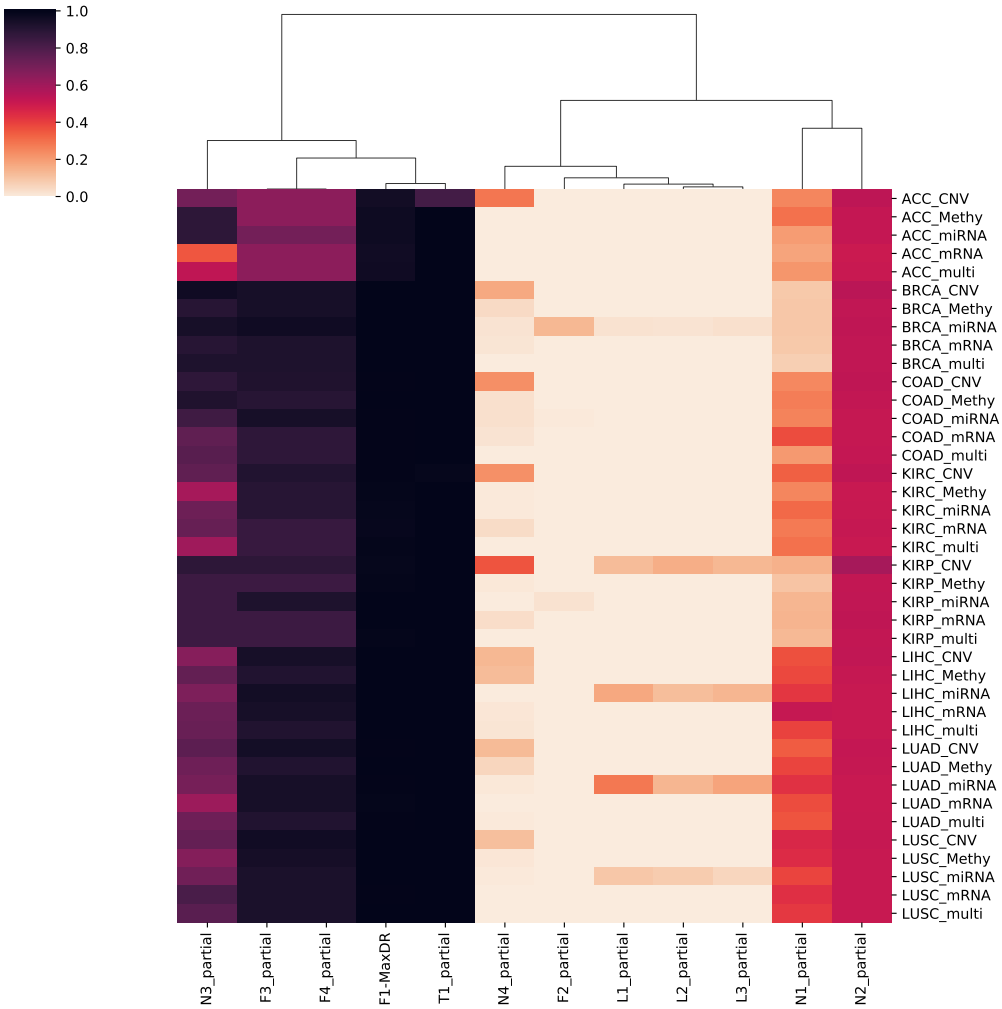
Fonte: O Autor

Figura A.93 – Clustermap Para a Classe No - Modelo Naive Bayes - Por Conjuntos de Dados - Complexidade Apenas



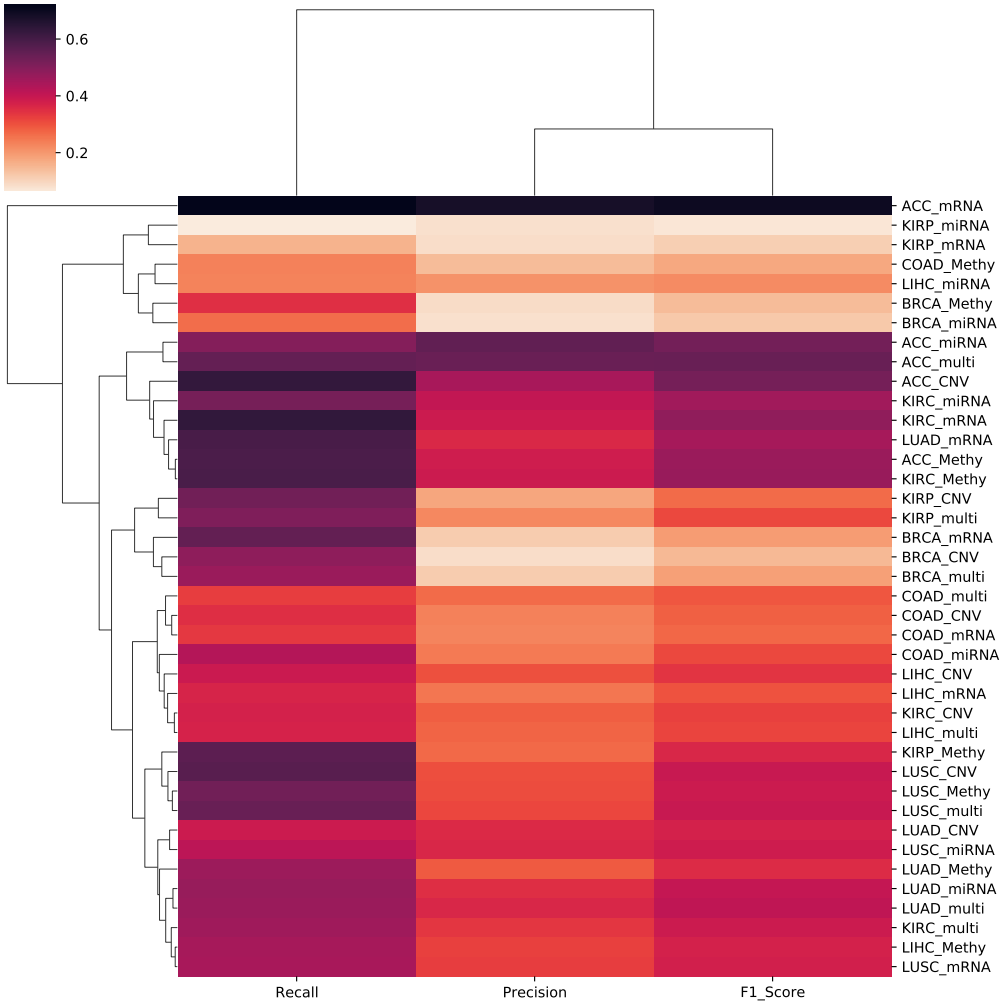
Fonte: O Autor

Figura A.94 – Clustermap Para a Classe No - Modelo Naive Bayes - Por Medida - Complexidade Apenas



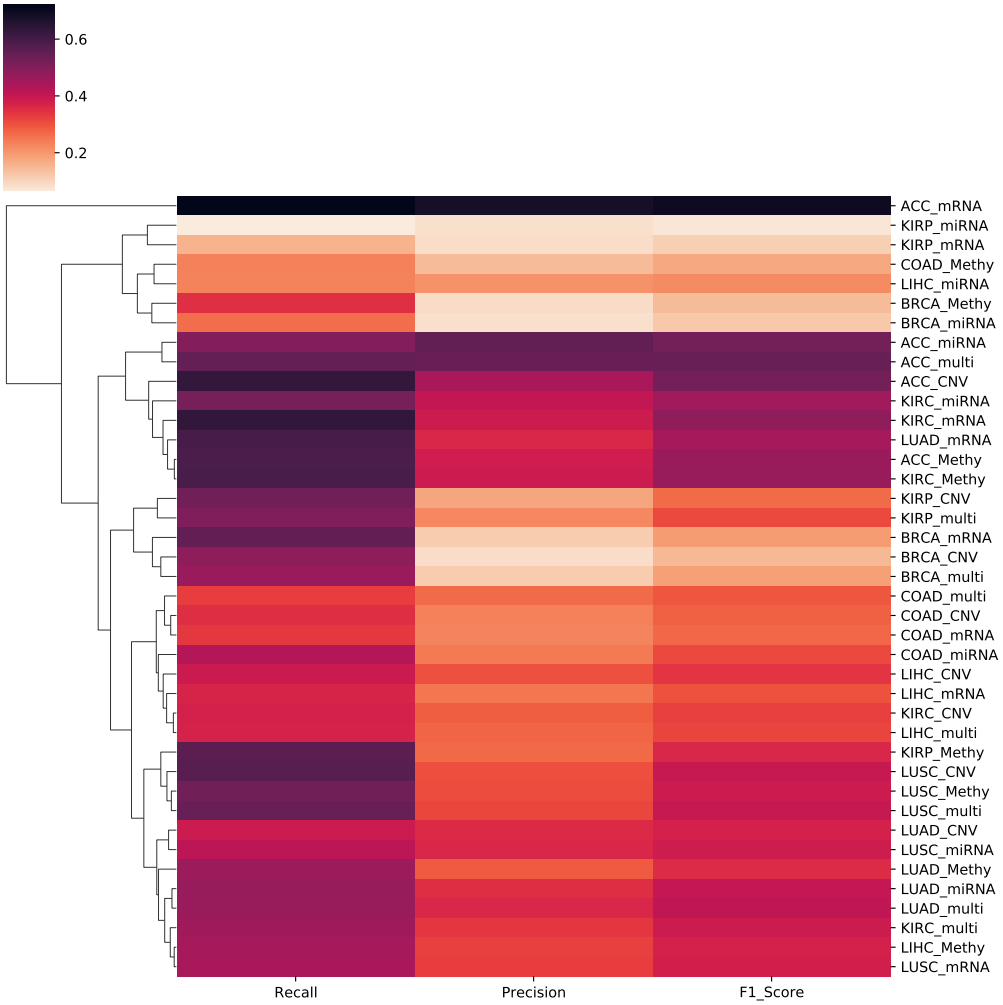
Fonte: O Autor

Figura A.95 – Clustermap Para a Classe No - Modelo Naive Bayes - Desempenho Apenas



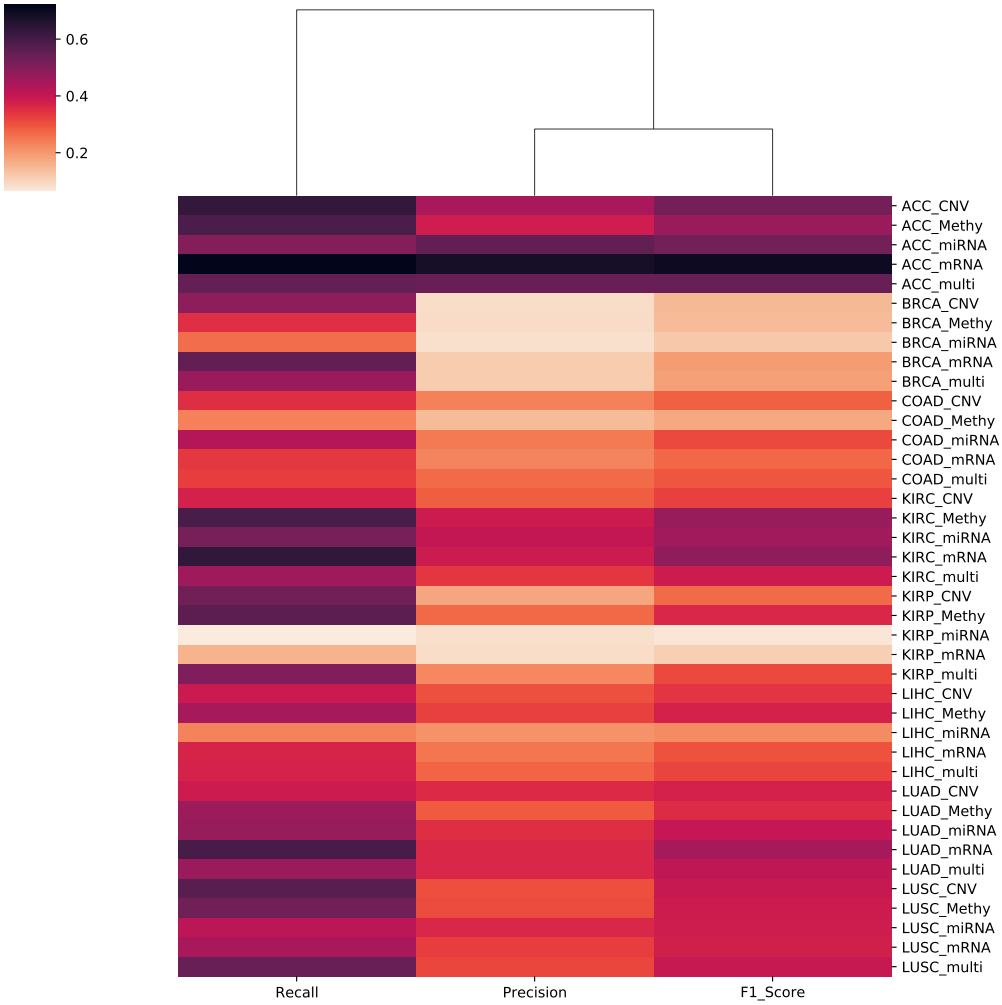
Fonte: O Autor

Figura A.96 – Clustermap Para a Classe No - Modelo Naive Bayes - Por Conjuntos de Dados - Desempenho Apenas



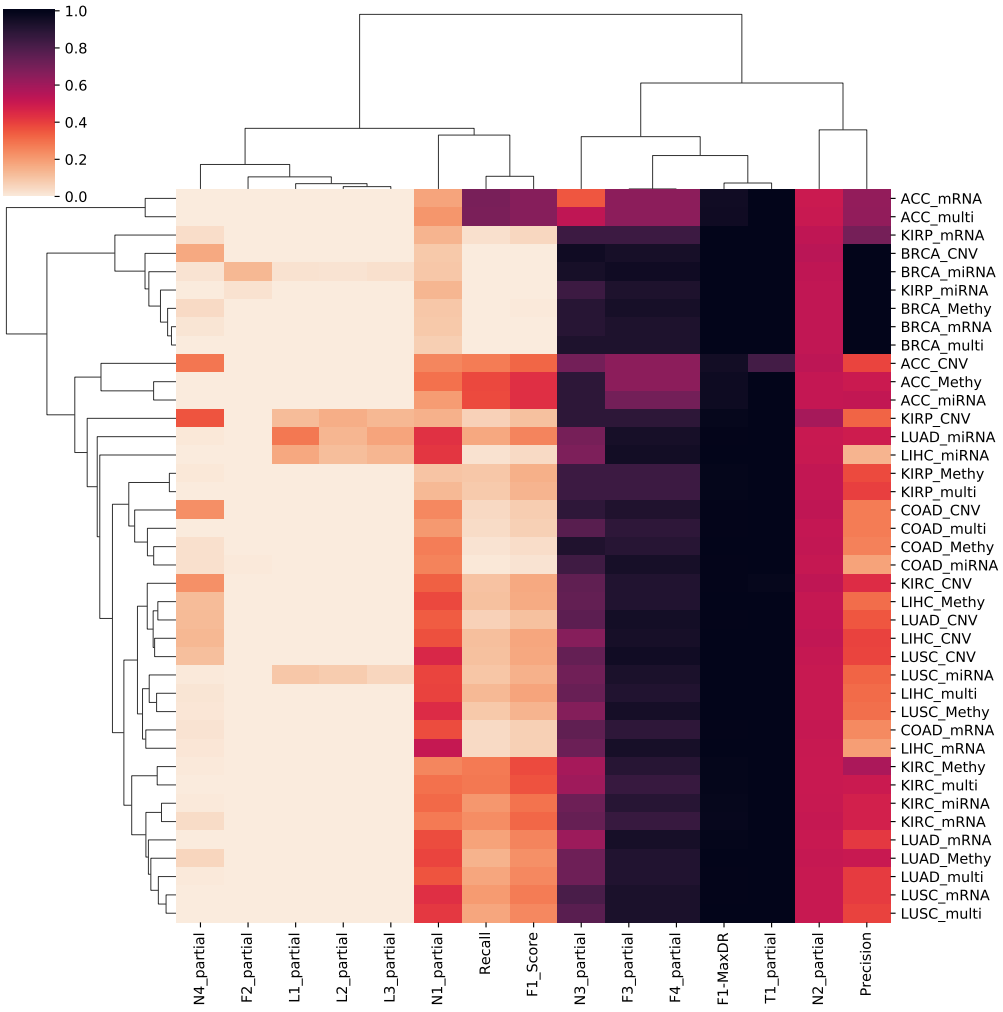
Fonte: O Autor

Figura A.97 – Clustermap Para a Classe No - Modelo Naive Bayes - Por Medida - Desempenho Apenas



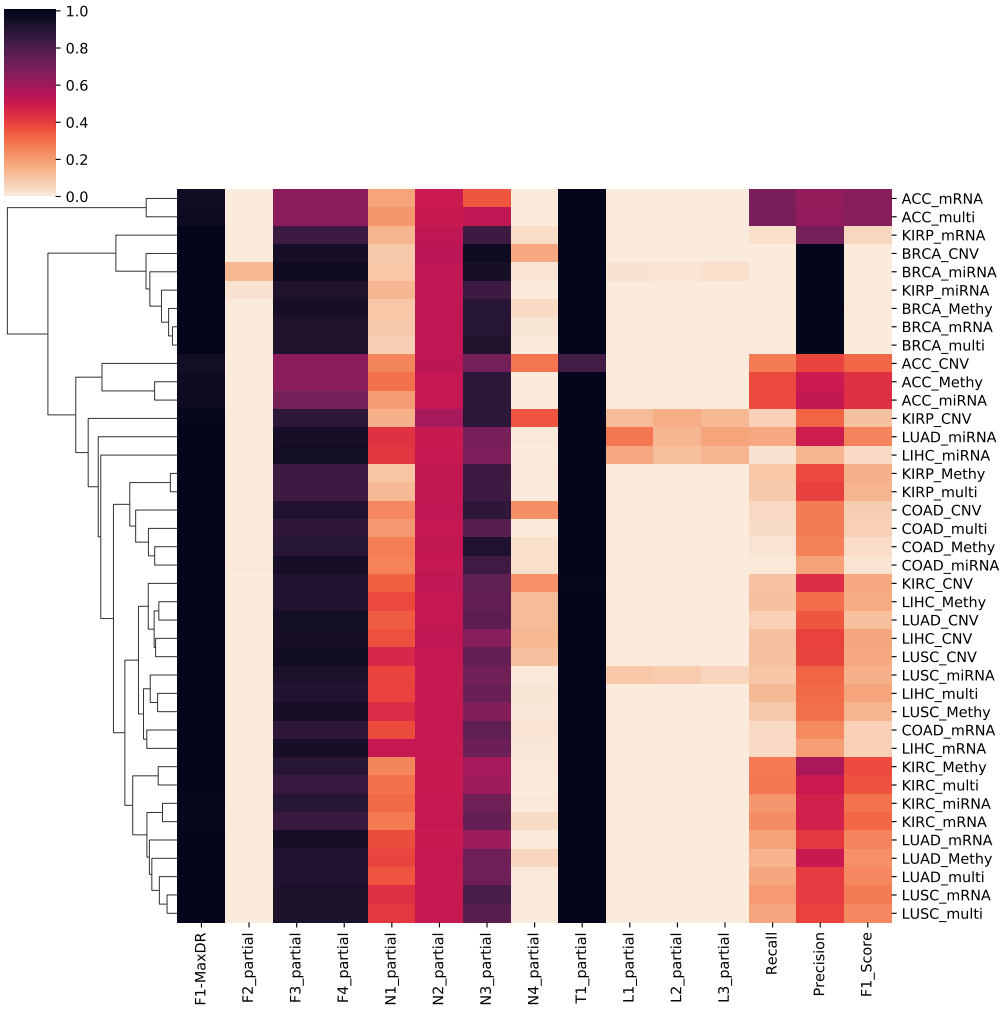
Fonte: O Autor

Figura A.98 – Clustermap Para a Classe No - Modelo GLM - Todas as Medidas



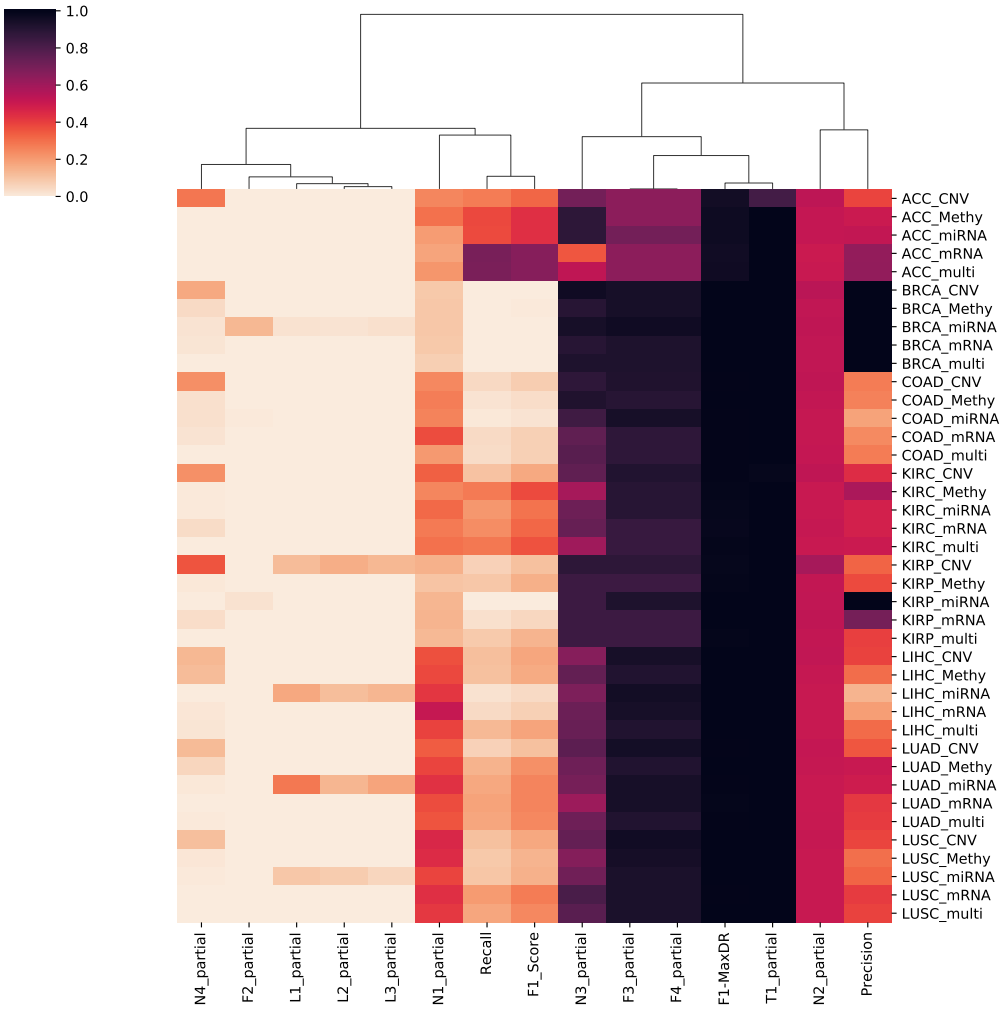
Fonte: O Autor

Figura A.99 – Clustermap Para a Classe No - Modelo GLM - Por Conjuntos de Dados - Todas Medidas



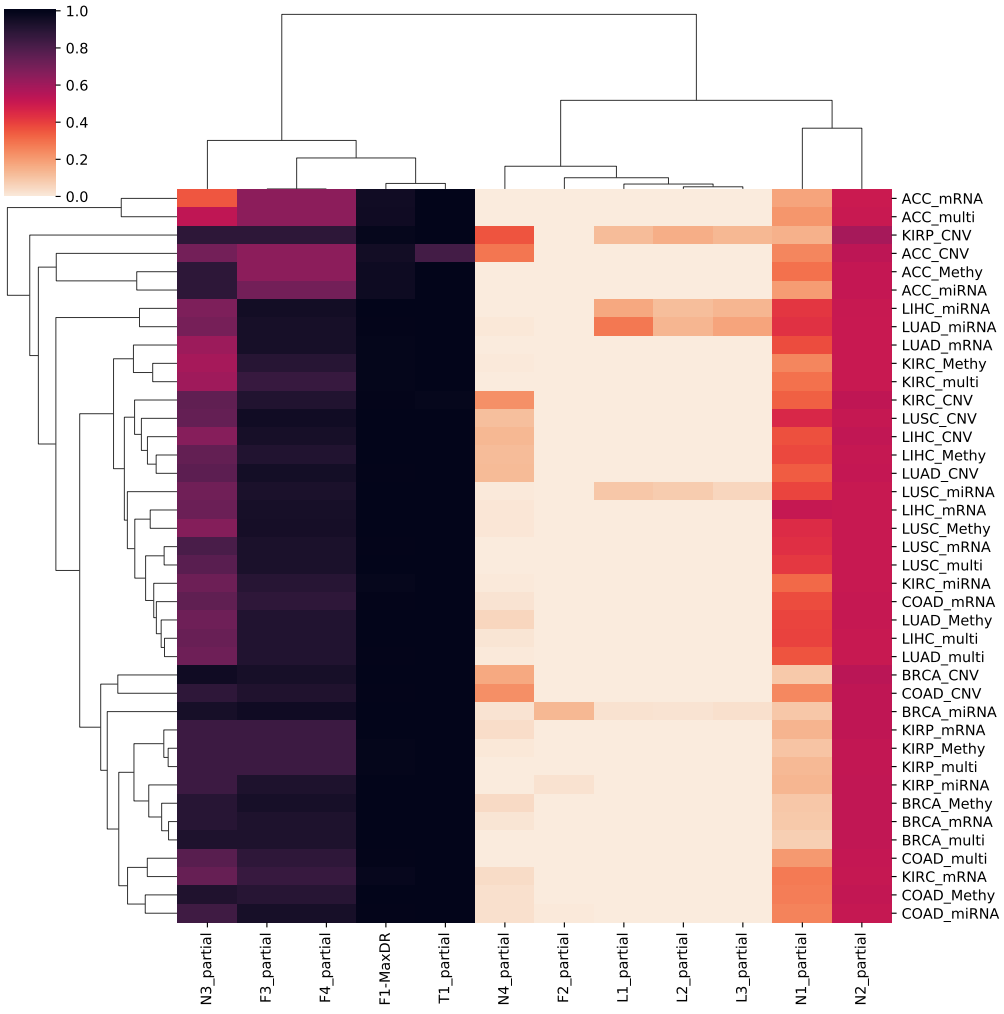
Fonte: O Autor

Figura A.100 – Clustermap Para a Classe No - Modelo GLM - Por Medida - Todas Medidas



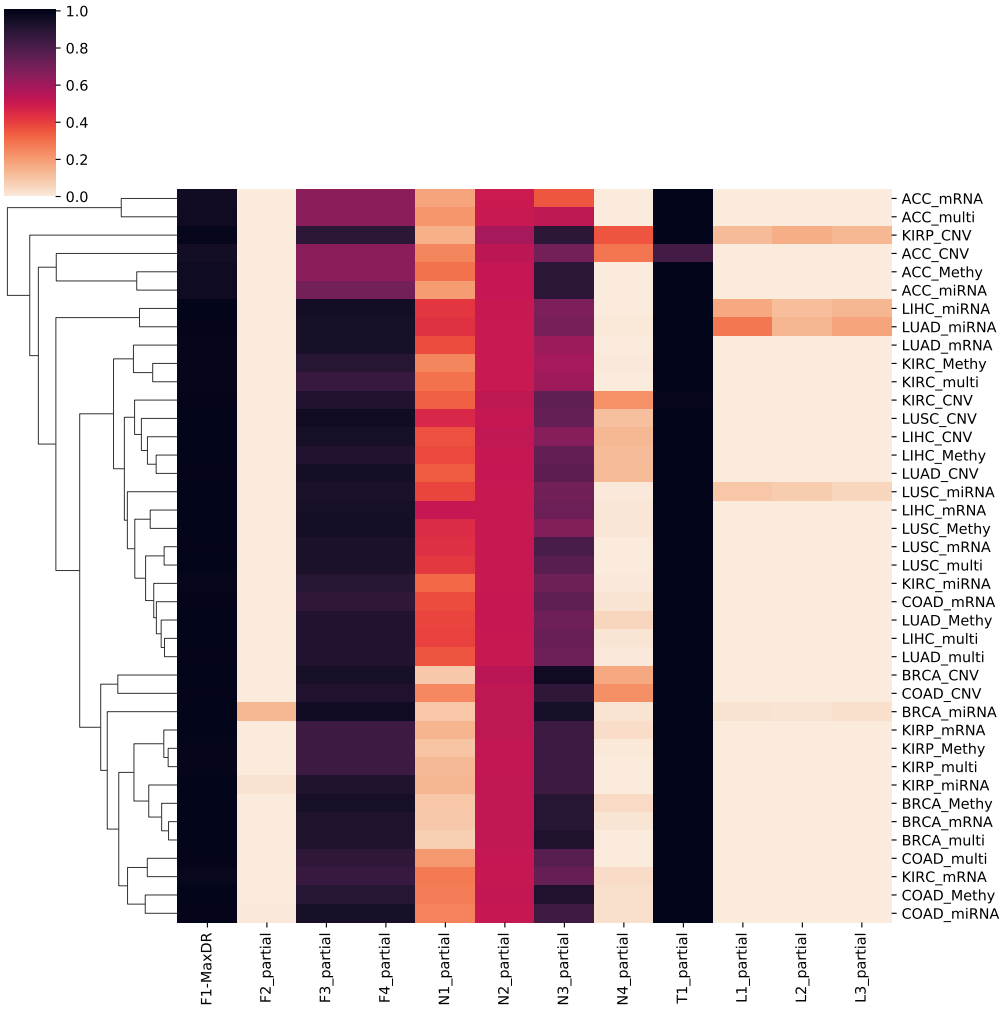
Fonte: O Autor

Figura A.101 – Clustermap Para a Classe No - Modelo GLM - Complexidade Apenas



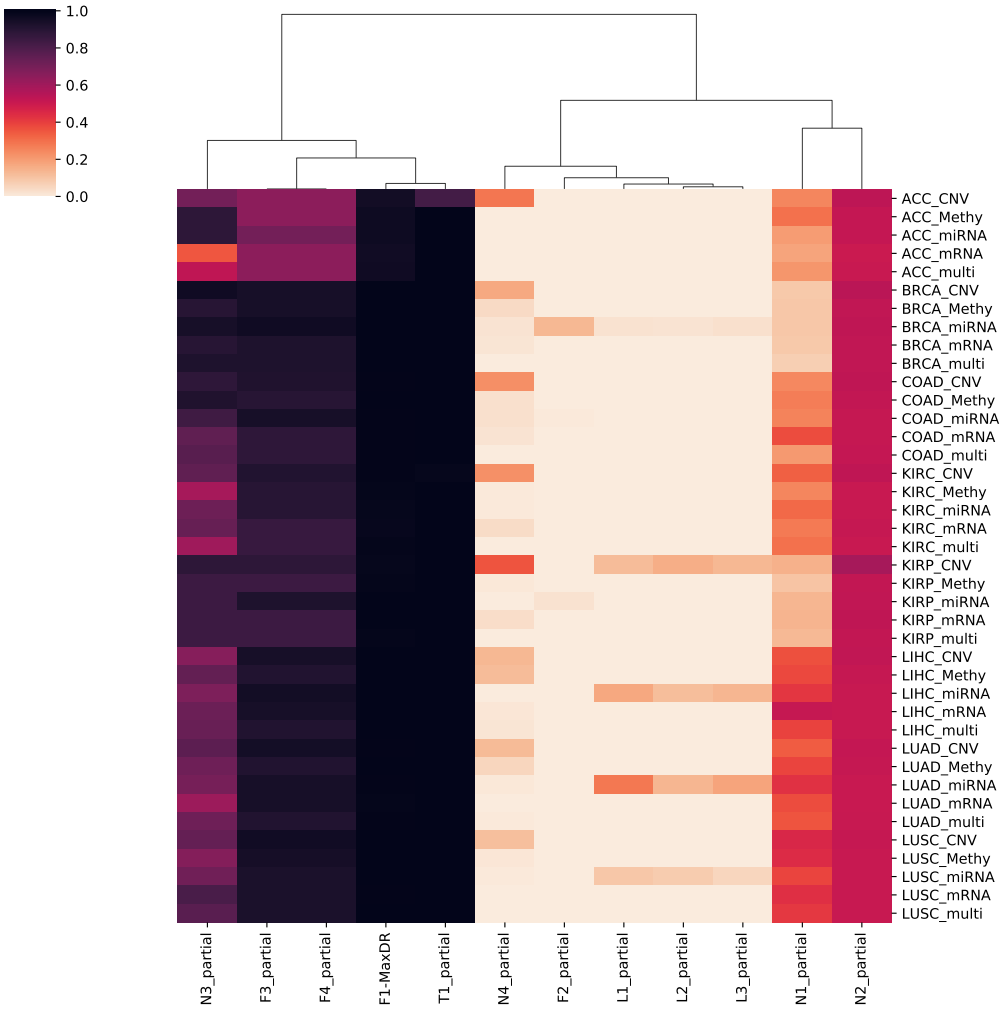
Fonte: O Autor

Figura A.102 – Clustermap Para a Classe No - GLM - Por Conjuntos de Dados - Complexidade Apenas



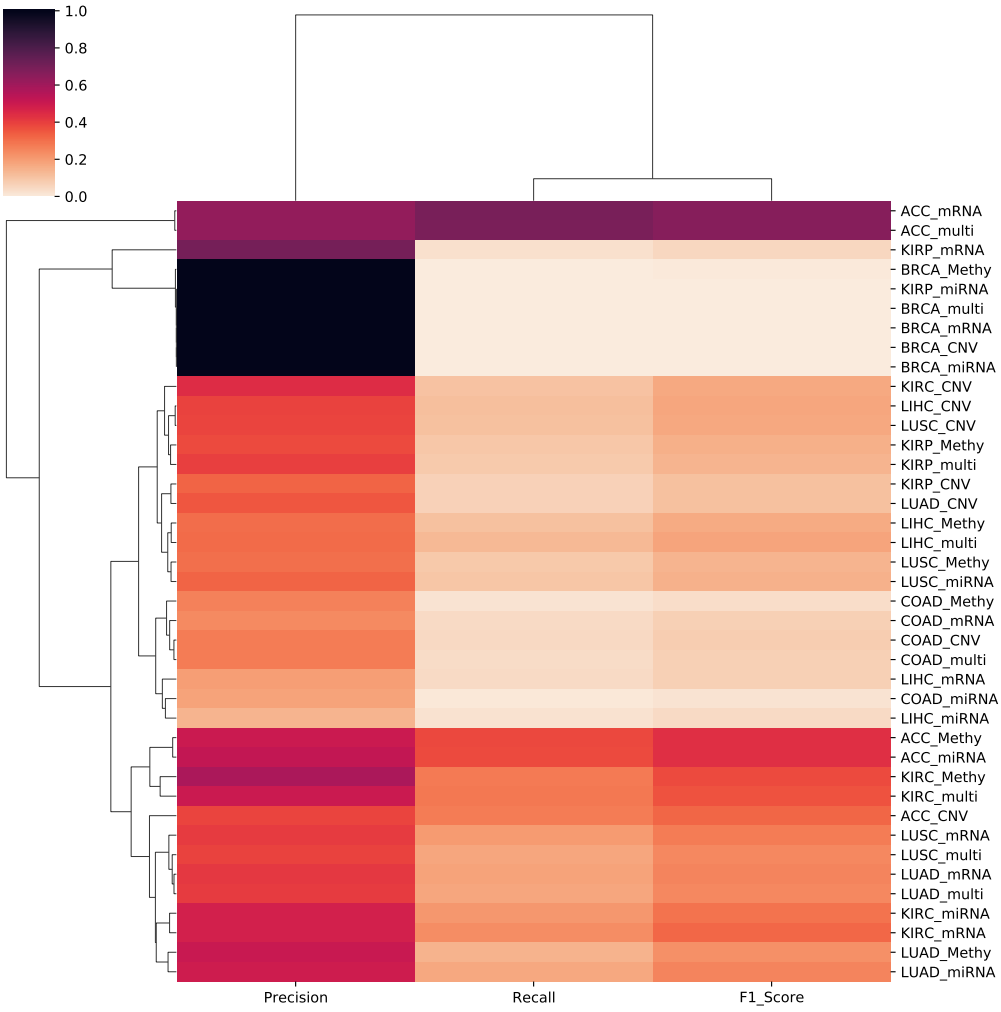
Fonte: O Autor

Figura A.103 – Clustermap Para a Classe No - GLM - Por Medida - Complexidade Apenas



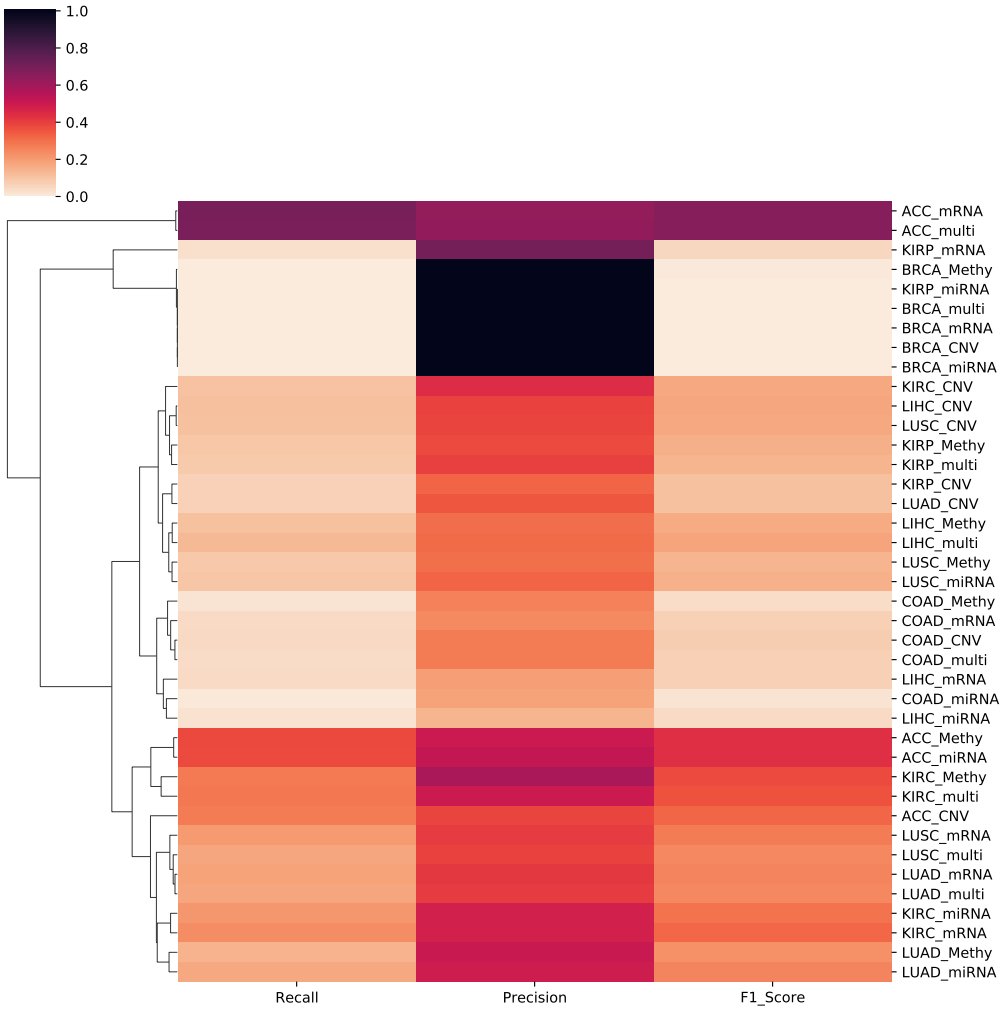
Fonte: O Autor

Figura A.104 – Clustermap Para a Classe No - Modelo GLM - Desempenho Apenas



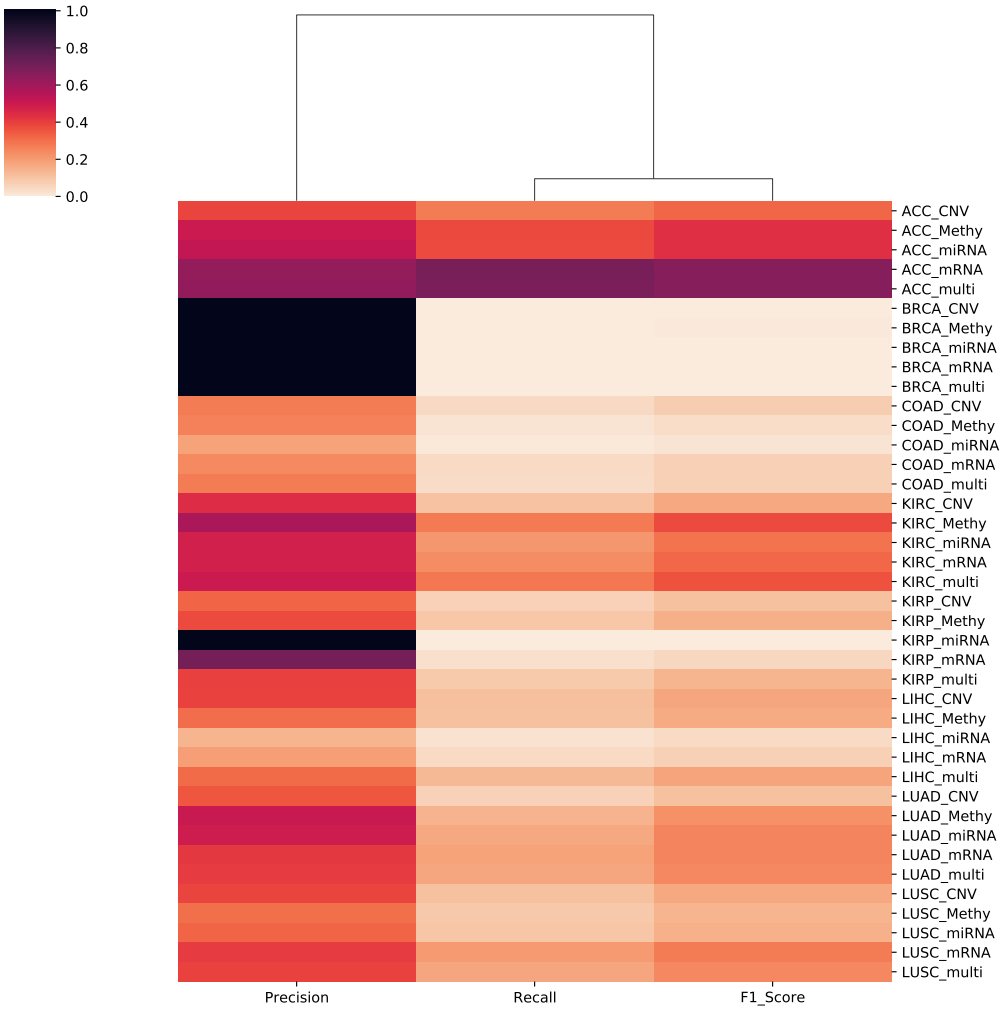
Fonte: O Autor

Figura A.105 – Clustermap Para a Classe No - Modelo GLM - Por Conjuntos de Dados - Desempenho Apenas



Fonte: O Autor

Figura A.106 – Clustermap Para a Classe No - Modelo GLM - Por Medida - Desempenho Apenas



Fonte: O Autor