GERAÇÃO E PROCESSAMENTO DE BANCOS DE DADOS DE EXPRESSÃO GÊNICA DE BIOINFORMÁTICA

Victor Ehiti Itimura Tamay (UTFPR) E-mail: victortamay@alunos.utfpr.edu.br
Glaucia Maria Bressan (UTFPR) E-mail: glauciabressan@utfpr.edu.br

Resumo: Inserido na interseção entre Bioinformática, Biologia Computacional e saúde pública, este estudo aborda o problema do pré-processamento de dados genômicos. Com a crescente de dados genômicos e expressão gênica, os métodos para a filtragem de dados têm se destacado como uma técnica fundamental para a organização de registros, extrair materiais e desenvolver novas hipóteses de pesquisa. Neste contexto de pesquisa, o objetivo deste trabalho é a geração de um pipeline automatizado, além de um conjunto de dados de expressão gênica, usando técnicas de processamento e tratamento de banco de dados. O pipeline foi aplicado em estudos relacionados à o estudo do Alzheimer e também para a classificação de subtipos de câncer de mama, demonstrando sua aplicabilidade em diferentes cenários biomédicos. Esse objetivo foi alcançado a partir do uso do Python e de suas respectivas bibliotecas, além do auxílio de repositórios públicos bem conceituados. Foi realizada a elaboração dos dados, por início da extração, fusão, organização das colunas e a normalização. Um dos principais destaques alcançados foi a criação de uma função específica para automatizar o fluxo de preparação dos dados, reduzindo a intervenção manual e aumentando a reprodutibilidade do processo. O resultado prático demonstra a construção de bases de dados padronizadas e consistentes, com potencial de reutilização científica. A validação das etapas demonstrou a eficácia da abordagem e reitera a importância crítica do pré-processamento para a obtenção de resultados confiáveis em Bioinformática.

Palavras-chave: Bioinformática; Pré-processamento de dados; Python; Expressão gênica.

GENERATION AND PROCESSING OF GENE EXPRESSION DATABASES IN BIOINFORMATICS

Abstract: Situated at the intersection of Bioinformatics, Computational Biology, and public health, this study addresses the challenge of genomic data preprocessing. With the increasing volume of genomic and gene expression data, data filtering methods have become a fundamental technique for organizing records, extracting relevant information, and supporting the development of new research hypotheses. In this research context, the objective of this work is the development of an automated pipeline, as well as the generation of a gene expression dataset, using data processing and database handling techniques. The pipeline was applied to studies related to Alzheimer's disease and to the classification of breast cancer subtypes, demonstrating its applicability across different biomedical scenarios. This objective was achieved through the use of Python and its associated libraries, along with data obtained from well-established public repositories. The methodology involved data preparation steps including extraction, integration, column organization, and normalization. One of the main highlights of this work is the implementation of a specific function to automate the data preparation workflow, reducing manual intervention and increasing process reproducibility. The practical outcome demonstrates the construction of standardized and consistent datasets with potential for scientific reuse. The validation of the steps confirmed the effectiveness of the proposed approach and reinforces the critical importance of preprocessing for obtaining reliable results in Bioinformatics.

Keywords: Bioinformatics; Data Preprocessing; Python; Gene Expression

1. Introdução

Recentemente, a integração entre Bioinformática, Biologia Computacional e saúde pública tem possibilitado evoluções significativas na compreensão e análise de dados

biológicos complexos de diversas naturezas, permitindo uma transição mais robusta da pesquisa de bancada para aplicações clínicas diretas (DAWADI et al., 2025; CURTIS et al, 2012). Essa convergência interdisciplinar tem permitido não apenas o desenvolvimento de novas abordagens analíticas, mas também a criação de ferramentas computacionais capazes de lidar com a crescente diversidade e volume de informações provenientes de diferentes áreas das ciências biológicas.

Devido ao grande avanço na quantidade de dados genômicos e de expressão gênica, destacam-se os métodos de mineração de dados (HAN et al, 2012; TAN et al, 2018; WITTEN et al, 2011), os quais têm se consolidado como instrumentos essenciais para a organização, filtragem, extração e análise de grandes volumes de dados, permitindo não apenas o gerenciamento eficiente de registros, mas também a identificação de padrões ocultos e complexos, correlações relevantes e geração de novas hipóteses no campo da Bioinformática, ampliando as possibilidades de investigação científica.

Os campos de Bioinformática e Biologia Computacional utilizam métodos de estudo de dados para a exploração e interpretação de um grande número de dados biológicos (ARNOLD et al, 2022). Tais estudos são fundamentais para interpretação de critérios em dados genômicos, modelagem de sistemas biológicos, especificação de desfechos e identificação de novas interseções entre dados genéticos (SMOLARZ et al., 2022). No entanto, a qualidade e precisão das análises depende diretamente da etapa de pré-processamento dos dados, a qual ainda representa um obstáculo significativo, especialmente no âmbito de dados de larga escala.

Não obstante dos avanços existentes, no contexto da literatura, é notório uma lacuna relacionada à padronização e automatização de pipelines de pré-processamento de dados genômicos, particularmente no que diz respeito à reprodutibilidade e escalabilidade das análises e à integração de ambientes de programação distintos (FANFANI et al., 2025; VAN et al., 2024). Muitos estudos se dedicam a aplicação de técnicas analíticas avançadas, porém não oferecem descrições detalhadas nem soluções generalizáveis para o pré-processamento dos dados, etapa crucial para a confiabilidade dos resultados obtidos.

Perante o exposto, este trabalho tem como objetivo a geração e o pré-processamento de bancos de dados de expressão gênica no campo da Bioinformática, utilizando a linguagem Python e suas extensões, sendo as principais delas o Pandas e Numpy, para a melhor manipulação de dados, para o aprofundamento no estudo da doença do Alzheimer e para a identificação e classificação de subtipos de câncer de mama, demonstrando a aplicabilidade do pipeline em diferentes contextos biomédicos.

A principal colaboração é o desenvolvimento de um fluxo automatizado para a construção novos dados de matrizes gênicas padronizadas, a partir de dados derivados de repositórios públicos, além de transformar o acesso público, promovendo maior acessibilidade, reprodutibilidade, potencial de reutilização científica e assim contribuindo para a evolução das pesquisas no campo da Bioinformática. Ademais, o trabalho pertencente ao grupo de Bioinformática da Universidade Tecnológica Federal do Paraná Campus Cornélio Procópio, tem contribuído para o desenvolvimento de trabalhos de conclusão de curso e também em dissertações de mestrado.

Assim, além de sua relevância para o avanço das técnicas de mineração de dados, este estudo se destaca por situar-se na interseção de diferentes campos do conhecimento, abordando de forma integrada contratempos relacionados ao pré-processamento de larga escala, potencializando a ocorrência de descobertas inovadoras. Nesse contexto, este trabalho contribui com a disponibilização de conjuntos de dados robustos e

fundamentados, capazes de fomentar o desenvolvimento de novas abordagens analíticas e de metodologias mais eficazes, bem como de assegurar a obtenção de resultados mais precisos e confiáveis em Bioinformática.

2. Materiais e métodos

Este estudo foi realizado por meio da utilização do ambiente de programação Google Colab, um serviço hospedado na nuvem baseado no Jupyter Notebook, que se destaca por não requerer instalação ou configuração prévia para uso, proporcionando acesso remoto e gratuito a recursos computacionais de alto desempenho, incluindo GPUs e TPUs. Essa infraestrutura é especialmente adequada para aplicações em aprendizado de máquina, ciência de dados e atividades educacionais, uma vez que permite a execução de código em nuvem com integração a diversas bibliotecas científicas.

A escolha da linguagem Python para o desenvolvimento das rotinas deve-se à sua ampla adoção pela comunidade científica e à disponibilidade de um ecossistema consolidado de bibliotecas voltadas à análise e manipulação de dados (CHOUDHARY et al., 2025; ANTAO, 2015). A biblioteca fundamental para este estudo é a chamada "Pandas (WES, 2013)", ferramenta central para a manipulação e análise de dados. A biblioteca Pandas oferece estruturas de dados eficientes, como DataFrames, que permitem realizar operações complexas de filtragem, agregação e transformação de maneira otimizada e com sintaxe intuitiva.

Os conjuntos de dados originais empregados foram obtidos de repositórios de alta credibilidade, incluindo AlzData[1], Kaggle[2], *Gene Expression Omnibus* (GEO)[3], NIAGADS[4], e *The Cancer Genome Atlas* (TCGA)[5]. Esses repositórios contêm dados em diferentes formatos, abrangendo desde arquivos tabulares (.csv, .tsv e .txt) até arquivos .json, imagens biomédicas e matrizes de expressão gênica. Como critérios para a seleção dos datasets, foi analisada a relevância do conjunto para o problema investigado, como a expressão gênica associada a condições específicas, assim como a disponibilidade pública e conformidade com diretrizes éticas, presença de metadados estruturados, como a identificação de amostras e genes, e também a qualidade e integridade das informações, como a ausência de inconsistências críticas.

Os dados selecionados são provenientes de técnicas como *microarray* ou RNA-seq (PEVSNER, 2015), refletindo o nível de expressão de cada gene em condições e contextos experimentais variados. Essa organização permite a aplicação de métodos estatísticos e de aprendizado de máquina para identificar padrões, relações e potenciais marcadores biológicos de relevância para a área de estudo.

Considerando que grande parte dos repositórios públicos renomados disponibiliza os conjuntos de dados no formato ".txt" (separado por tabulações), desenvolveu-se uma função específica para o tratamento e conversão desses arquivos para o formato .csv, amplamente compatível com ferramentas analíticas e pipelines de aprendizado de máquina. Essa função é intitulada "toCSV". Alternativamente de materiais convencionais de conversão, essa função proposta incorpora uma leitura parametrizada de delimitadores, o tratamento de exceções para arquivos inconsistentes ou corrompidos, a padronização automática de cabeçalhos e a verificação básica de integridade, como o número de colunas por linhas.

Em primeiro lugar, é necessário adicionar o arquivo original a ser tratado ao ambiente de execução, nesse caso é o `TC_GSE15222_exp.txt`. A partir disso também se faz necessário nomear o novo arquivo que será gerado no formato .csv, então

`TC_GSE15222_exp.csv`. Feito isso, a função `pd.read_csv(arquivo_txt, delimiter="\t")` irá ler o arquivo de texto, e para cada `\t` ou "`Tab`", ele irá dar início à uma nova coluna. O resultado desse processo é a formação do arquivo agora no formato `.csv`. Além da conversão, a função incorpora mecanismos básicos de tratamento de exceções para lidar com eventuais erros de leitura, garantindo maior robustez e reprodutibilidade do processo. O Algoritmo 1 apresenta a função desenvolvida para a coleta e geração de conjuntos de dados.

```python
import pandas as pd

arquivo_txt = "TC_GSE15222_exp.txt"
arquivo_csv = "TC_GSE15222_exp.csv"

try:
    df = pd.read_csv(arquivo_txt, delimiter="\t")

    df.to_csv(arquivo_csv, index=False)

    print("\nArquivo convertido com sucesso!")

except FileNotFoundError:
    print(f"\nERRO: O arquivo '{arquivo_txt}' não foi encontrado.")

except Exception as e:
    print(f"\nERRO: Ocorreu um problema inesperado durante a
conversão.")
    print(f"Detalhe do erro: {e}")
```

Algoritmo 1 – Função toCSV

1 http://www.alzdata.org/.
2 https://www.kaggle.com/.
3 https://www.ncbi.nlm.nih.gov/geo/.
4 https://www.niagads.org/.
5 https://www.cancer.gov/ccg/research/genome-sequencing/tcga.

O repositório NIAGADS (*National Institute of Aging Genetics of Alzheimer's Disease Data Storage Site*) é um grande repositório que armazena dados genômicos da doença de Alzheimer. O AlzData, desenvolvido pelo mesmo grupo, constitui uma interface moderna e interativa que facilita a exploração e análise desses dados. Embora tais repositórios disponibilizem amplos volumes de informações de forma pública, determinadas bases contendo dados genômicos humanos detalhados permanecem sob acesso restrito, devido a exigências éticas e legais de proteção de dados sensíveis.

Para a obtenção de dados restritos, torna-se necessária a submissão de um (*Data Access Request*), incluindo justificativa científica e declaração de conformidade ética, a ser avaliada por um comitê especializado. Os dados públicos, que foram utilizados neste estudo, são reais, permitindo que pesquisadores de todo o mundo utilizem esses dados

como fonte de informações para avanços nas pesquisas de diversas áreas.

Após a seleção do conjunto de dados mais adequado aos objetivos da investigação, o arquivo é processado pela função descrita no Algoritmo 1, que converte e organiza o conteúdo para o formato .csv. A etapa seguinte consiste na análise exploratória inicial, na qual a primeira linha do arquivo é interpretada como cabeçalho, contendo os identificadores das colunas — tipicamente, nomes de genes e códigos de amostras (GSMs). A primeira coluna representa a lista de genes, enquanto as demais colunas contêm os valores de expressão gênica correspondentes às diferentes amostras experimentais. Assim, possibilitando a aplicação de métodos estatísticos e técnicas de aprendizado de máquina para a identificação de padrões e potenciais biomarcadores.

Apesar dos proveitos do método utilizado, há algumas limitações a serem consideradas, como a não preservação integral de estruturas complexas presentes em formatos como .json, assim como a possibilidade da utilização dos dados públicos tenham vieses associados à sua origem. Além disso, a função toCSV não inclui a normalização dos dados, tendo a necessidade da utilização de procedimentos adicionais para o adequado tratamento dos dados.

3. Resultados e discussão

Os resultados alcançados por esse estudo são fundamentais para demonstrar o destaque do processo de mineração, limpeza, tratamento e preparação dos dados para o estudo de dados de expressão gênica referentes aos subtipos de câncer de mama e Alzheimer. Para este estudo, a qualidade e a consistência dos dados utilizados, provenientes dos repositórios públicos renomados, foram fundamentais para a robustez das análises subsequentes e para a obtenção de informações diversificadas, abrangendo diferentes perfis de amostras e características genéticas. Contudo, é necessário mencionar sobre a avaliação de forma objetiva dos impactos obtidos nas etapas que foram diretamente relacionados na qualidade dos dados e na confiabilidade das análises.

Com o intuito de assegurar a consolidação e a reprodutibilidade das bases de dados geradas, foi empregada uma metodologia estruturada que contemplou múltiplas etapas. Inicialmente, realizou-se uma pesquisa sistemática e criteriosa em repositórios públicos de dados biológicos, priorizando aqueles com histórico de curadoria científica e ampla utilização pela comunidade acadêmica. Em seguida, foi desenvolvida, em Python, uma função dedicada ao tratamento e padronização dos conjuntos de dados obtidos, viabilizando a construção de um banco de dados estruturado, coerente e operacionalmente eficiente. Essa base consolidada, processada com o auxílio de ferramentas de mineração de dados, constitui-se como um recurso de alto valor para investigações futuras, contribuindo para o avanço da pesquisa biomédica.

Após a etapa de mineração, é necessário aplicar rigorosamente as técnicas de filtros no conjunto de dados para realizar a limpeza e tratamento, como: a remoção de informações duplicadas (evitando redundância e enviesamento das análises), a correção de dados nulos ou inconsistentes; e a normalização de dados, para harmonizar escalas e padrões, favorecendo comparações adequadas entre diferentes amostras e estudos, assegurando consistência e precisão (PEREIRA, 2021). Como forma de avaliação quantitativa do impacto dessa etapa, observou-se a redução de inconsistências estruturais, a padronização dos intervalos de valores de expressão gênica, além da melhoria na distribuição estatística dos dados, como a redução de variabilidade não biológica.

Esse processo é fundamental para garantir o bom funcionamento e precisão da análise.

Adicionalmente, a transformação de alguns conjuntos de dados no formato ".txt" para o formato ".csv" por meio da função desenvolvida facilitou a manipulação e análise subsequente, garantindo que os dados estivessem prontos para serem utilizados de maneira eficiente e eficaz.

Para os conjuntos de dados de Alzheimer obtidos, ordenados por linhas e colunas, foi verificado uma estrutura consistente após o pré-processamento, como as colunas se ordenaram como: Gene, onde cada entrada nesta coluna é o nome de um gene específico, servindo como um identificador de cada linha; A segunda coluna, representa amostras individuais, ou seja, são identificadores padrão para amostras do repositório de dados de expressão gênica GEO, cada amostra podendo ser um paciente, um tecido biológico ou uma condição experimental diferente, usualmente iniciando como "GSM"; O restante das colunas nomeadas por números de pontos flutuantes, representam o nível de expressão de um determinado gene em uma amostra específica. Passando para análise das linhas, cada uma correspondente a um único gene, fornecendo informações sobre esse gene em diferentes amostras analisadas.

Utilizando o conjunto de dados HP_GSE29378 de exemplo para explicação, que tem sido amplamente utilizado para o estudo de genes e redes de co-expressão na doença do Alzheimer (CHEN et al., 2024), possuindo 19 901 genes (linhas) e 32 amostras (colunas nomeadas por GSM...), totalizando 636.832 pontos de dados de expressão, com um valor máximo de expressão de 15.07; valor mínimo de 6.57; uma média de 8.40; e por último um desvio padrão de 1.49, evidenciando uma distribuição moderadamente concentrada dos dados em torno da média obtida, refletindo uma boa consistência após o processo.

No que se refere ao dados de câncer de mama (TCGA-BRCA), O conjunto utilizado é um arquivo .zip que conta com quatro arquivos .csv (TCGA-BRCA-RNA-Seq; TCGA-BRCA-A2-target_variable; TCGA-BRCAA2-CLINI e TCGA-BRCA-A2_SLIDE) e mais um .zip com as imagens (TCGA-BRCA-A2-DEEPMED-TILES). É notório a integração de diversas fontes de dados, incluindo informações de expressão gênica, variáveis clínicas, variáveis-alvo e imagens histopatológicas, ampliando o potencial analítico, em contra-mão , impondo desafios adicionais para a consistência das informações.

O arquivo TCGA-BRCA-RNA-Seq, com estrutura tabular, com genes organizados nas linhas e amostras nas colunas, conta com aproximadamente 20 mil genes e 1.200 amostras, enquanto o TCGA-BRCA-A2-CLINI representa informações clínicas e genômicas detalhadas dos pacientes, incluindo 219 variáveis, como dados demográficos, características tumorais, marcadores biológicos e informações de sobrevida.

O dataset TCGA-BRCA-A2-target_variable define o objetivo de uma tarefa de classificação, essencial para um modelo de aprendizado de máquina, no qual o resultado esperado seria prever se o tumor é positivo ou negativo. Possuindo apenas duas colunas, sendo a principal informação o ER_STATUS, fundamental para orientar decisões de tratamento. O conjunto de dados conta com 586 amostras no total, sendo 449 amostras positivas e 137 negativas.

A Figura 1 clarifica essa distribuição e além do seu fundamento descritivo, é evidenciado um desbalanceamento significativo entre as classes. Ainda assim, há a possibilidade desse cenário acarretar no favorecimento da classe majoritária por conta do modelo preditivo, como é observado em datasets oncológicos de larga escala, exigindo o uso de modelos interpretáveis e métodos de aprendizado de máquina de alta

eficiência para mitigar o viés da classe majoritária (CHOWDHURY et al., 2025).

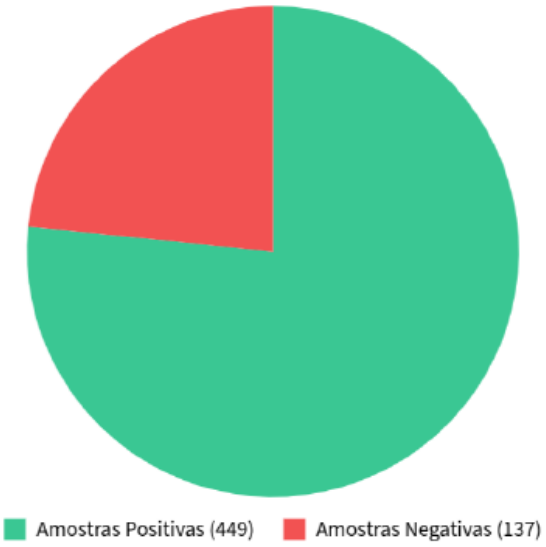

Figura 1 – Gráfico com análise de amostras positivas e negativas

Por fim, o TCGA-BRCA-A2_SLIDE tem a função de integrar os dados clínicos aos arquivos de imagem, contendo apenas as colunas PATIENT, Sample ID, Slide ID, File Name e File ID. O arquivo mapeia um total de 1096 imagens. No entanto, devido a quantidade de subpastas dentro do arquivo, há os desafios relacionados à sincronização das informações.

Em suma, os resultados obtidos evidenciam a conexão direta entre o processo de preparação dos dados e a qualidade das análises, não apenas com intuito da viabilização da utilização dos dados, mas também para a garantia da confiabilidade e validação dos resultados.

4. Conclusão

De acordo com os resultados obtidos, o processo de mineração, limpeza, filtragem, tratamento e preparação dos dados foi fundamental para a conquista de um conjunto de dados conciso e preciso. Isso mostra o destaque de um conjunto de dados de boa qualidade, o que reforça ainda mais a importância de hábitos rigorosos de pré-processamento.

Ademais, corrobora-se que cada etapa do processo metodológico foi essencial para garantir a integridade e consistência das informações, reforçando a solidez dos resultados apresentados. Como principal contribuição científica, o estudo demonstra a eficácia de um pipeline bem estruturado de aquisição e pré-processamento de dados aplicado a obstáculos de classificação e identificação de padrões em doenças relacionadas à expressão gênica, proporcionando uma base metodológica replicável para pesquisas futuras.

Todavia, há limitações a serem ponderadas. A dependência da qualidade das bases de dados utilizadas, assim como possíveis restringimentos quanto à diversidade amostral, é capaz de impactar a generalização dos resultados. Aditivamente, a ausência de validação em múltiplos contextos ou com diferentes configurações algorítmicas pode limitar a abrangência das conclusões.

Diante do exposto, como direções para trabalhos futuros, aconselha-se a ampliação das bases de dados, assim como a inclusão de novos algoritmos, como redes neurais, para o aperfeiçoamento da robustez metodológicas e à confiabilidade dos resultados, além da realização de validações cruzadas em múltiplos cenários. Ademais, a incorporação de dados multimodais pode acrescentar a capacidade preditiva dos modelos.

Desta forma, é evidente os avanços concretos na aplicação de técnicas de aprendizado de máquina à análise de expressão gênica, ressaltando os resultados obtidos, aliados às métricas apresentadas. Assim, contribuindo para o desenvolvimento de soluções mais robustas e confiáveis na área.

Referências

ANTAO, T. *Bioinformatics with Python cookbook*. Packt Publishing, 2015.

ARNOLD, M. *et al*. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, vol. 66, p. 15–23, 2022.

CHEN, Y.; LI, Z.; GE, X. *et al*. Identification of novel hub genes for Alzheimer's disease associated with the hippocampus using WGCNA and differential gene analysis. Frontiers in Neuroscience, v. 18, 1359631, 2024. DOI: 10.3389/fnins.2024.1359631.

CHOUDHARY, S. *et al.* Bridging the gap between R and Python in bulk transcriptomic data analysis. Nature Scientific Reports, 2025. DOI: 10.1038/s41598-025-03376-y.

CHOWDHURY, T. M. *et al*. An Efficient and Interpretable Machine Learning Model for Classifying Breast Cancer Subtypes Using Gene Expression Profiles. Engineering, Technology & Applied Science Research, v. 15, n. 1, p. 19283-19289, 2025. DOI: 10.48084/etasr.11179.

CURTIS, C. *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, Nature Publishing Group UK London, vol. 486, n.º 7403, p. 346–352, 2012.

DAWADI, P.; POKHAREL, B.; SHRESTHA, A. *et al*. From bench to bytes: a practical guide to RNA sequencing data analysis. Frontiers in Genetics, v. 16, 1697922, 2025. DOI: 10.3389/fgene.2025.1697922.

FANFANI, V. *et al*. Reproducible processing of TCGA regulatory networks. GigaScience, v. 14, giaf126, 2025. DOI: 10.1093/gigascience/giaf126.

HAN, J., KAMBER, M. & PEI, J.. *Data Mining: Concepts and Techniques*. 3. ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

PEREIRA, P. C.. *Introdução a Banco de Dados*. São Paulo: Editora Senac, 2021.

PEVSNER, J.. *Bioinformatics and functional genomics*. John Wiley & Sons, 2015.

SMOLARZ, B., NOWAK, A. Z. & ROMANOWICZ, H.. Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). *Cancers*, MDPI, vol. 14, p. 1–27, 2022.

TAN, P.N., STEINBACH, M. & KUMAR, V.. *Introduction to Data Mining*. 2. ed. USA: Pearson Education Limited, 2018.

VAN, R. *et al.* A comparison of RNA-Seq data preprocessing pipelines for transcriptomic predictions across independent studies. BMC Bioinformatics, v. 25, 181, 2024. DOI: 10.1186/s12859-024-05801-x.

MCKINNEY, W. *Python for data analysis*. O'Reilly Publishing, 2013.

WITTEN, I. H., FRANK, E. & HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed. Burlington, MA, USA: Morgan Kaufmann, 2011.