**Caracterização da Depressão em Adultos de Meia-Idade Utilizando Machine Learning e o Método CAPTO**

Pedro Henrique Rodrigues da Silva†  
Pontifícia Universidade Católica de Minas Gerais

Betim, Minas Gerais, Brasil  
1429338@sga.pucminas.br

**ABSTRACT**

Coloque aqui o resumo do trabalho. Inicie contextualizando o problema, fale os objetivos, a metodologia utilizada e os resultaods obtidos. É um resumo objetivo, mas que precisa contra toda a história realizada.

**KEYWORDS**

Coloque as palavras chaves aqui.

**ACM Reference format:**

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* <https://doi.org/10.1145/1234567890>

1**Introdução**

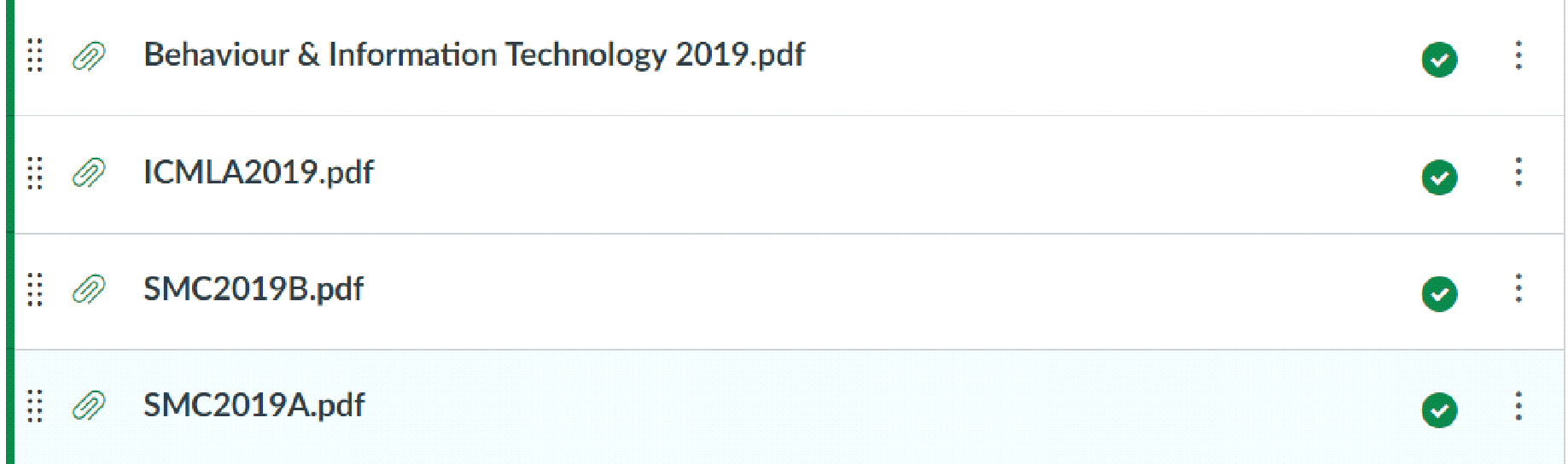
Faça uma pequena introdução, falando do contexto do trabalho e quais os seus objetivos.

O seu problema é sobre cancer? Fale um pouco sobre câncer. Use referências bibliográficas..

Lembre-se: a contextualização do problema é importante para a compreensão do que será lido depois.

Tem dúvida sobre isso: veja os artigos que coloquei no CANVAS para vocês se basearem. A Figura 1 mostra estes artigos que estão no CANVAS.

**Figura 1 – Artigos no CANVAS para vocês se basearem**



Estes artigos ajudarão vocês a fazerem o restante do trabalho também. Então, deem uma olhada neles com carinho.

**2. Descrição da base de dados**

A pesquisa conduzida em 2019 pela Pesquisa Nacional de Saúde é a fonte primária deste estudo, fornecendo uma rica coleção de dados relacionados à saúde, comportamentos e condições socioeconômicas. Esses dados abrangem uma amostra representativa da população, oferecendo insights valiosos para análises profundas e abrangentes.

Descrição Detalhada da Base de Dados:

Compreendendo a dimensão e a complexidade da pesquisa, é crucial destacar a meticulosa seleção e preparação dos dados. A base de dados original apresenta uma amplitude considerável, com 293.726 registros e 1.087 atributos. Utilizando uma abordagem cuidadosa de pré-processamento, empregamos o método CAPTO para reduzir essa vasta extensão para 40.359 entradas e 138 atributos relevantes para nossa investigação.

Dimensão: Características do Indivíduo

Nesta categoria, encontram-se detalhes demográficos e informações pessoais dos participantes:

- Sexo (C006): Define a categoria de gênero do indivíduo.

- Data de Nascimento (C00702 e C00703): Registrando mês e ano de nascimento para cálculos de idade.

- Idade (C008): Reflete a idade do respondente na data de referência da pesquisa.

Dimensão: Fatores Sociais e Ambientais

Essa seção destina-se a capturar elementos socioeconômicos e ambientais influentes na saúde:

- Rede de Apoio Social (M01401 e M01501): Avalia o suporte familiar e de amigos em momentos críticos.

- Rendimento e Receita (E0160X, E0180X): Detalha a origem e natureza dos ganhos financeiros dos participantes.

- Acesso a Recursos de Saúde (F001021, F007021, F008021): Indicações de benefícios previdenciários, pensões, aluguéis, entre outros.

Dimensão: Hábitos de Saúde

Este segmento concentra-se nos comportamentos e práticas relacionados à saúde:

- Atividade Física (P034, P035, P03701, P03702): Informa sobre a frequência e duração da prática de exercícios físicos.

- Consumo de Álcool (P027, P02801, P029): Avalia os padrões de ingestão de bebidas alcoólicas.

- Padrões Alimentares (P0060X, P00901, P01001, P01101, P013, P015, P018, P019, P0200X, P0210X, P023, P02401, P02501, P02602, P02601): Detalha a frequência e tipos de alimentos consumidos.

Dimensão: Antropometria

Essa dimensão apresenta informações antropométricas cruciais, como altura e peso:

- Altura e Peso (P00102, P00402, P00403, P00103, P00104, W00101, W00102, W00201, W00202): Fornecem dados precisos sobre as medidas físicas dos participantes.

- Índices de Massa Corporal (IMC): Calculados a partir das medidas de altura e peso, fornecendo uma visão da saúde física dos entrevistados.

Outras Dimensões

Ademais, a base de dados contém uma ampla gama de informações sobre diagnósticos médicos, tratamentos de saúde, condições habitacionais, acesso a serviços de saúde, entre outros aspectos relevantes para o estudo e análise aprofundada.

Essa descrição abrangente da estrutura da base de dados, suas dimensões e conteúdos, proporciona uma base sólida para a condução de análises robustas e a extração de insights significativos para o avanço do conhecimento na área da saúde pública.

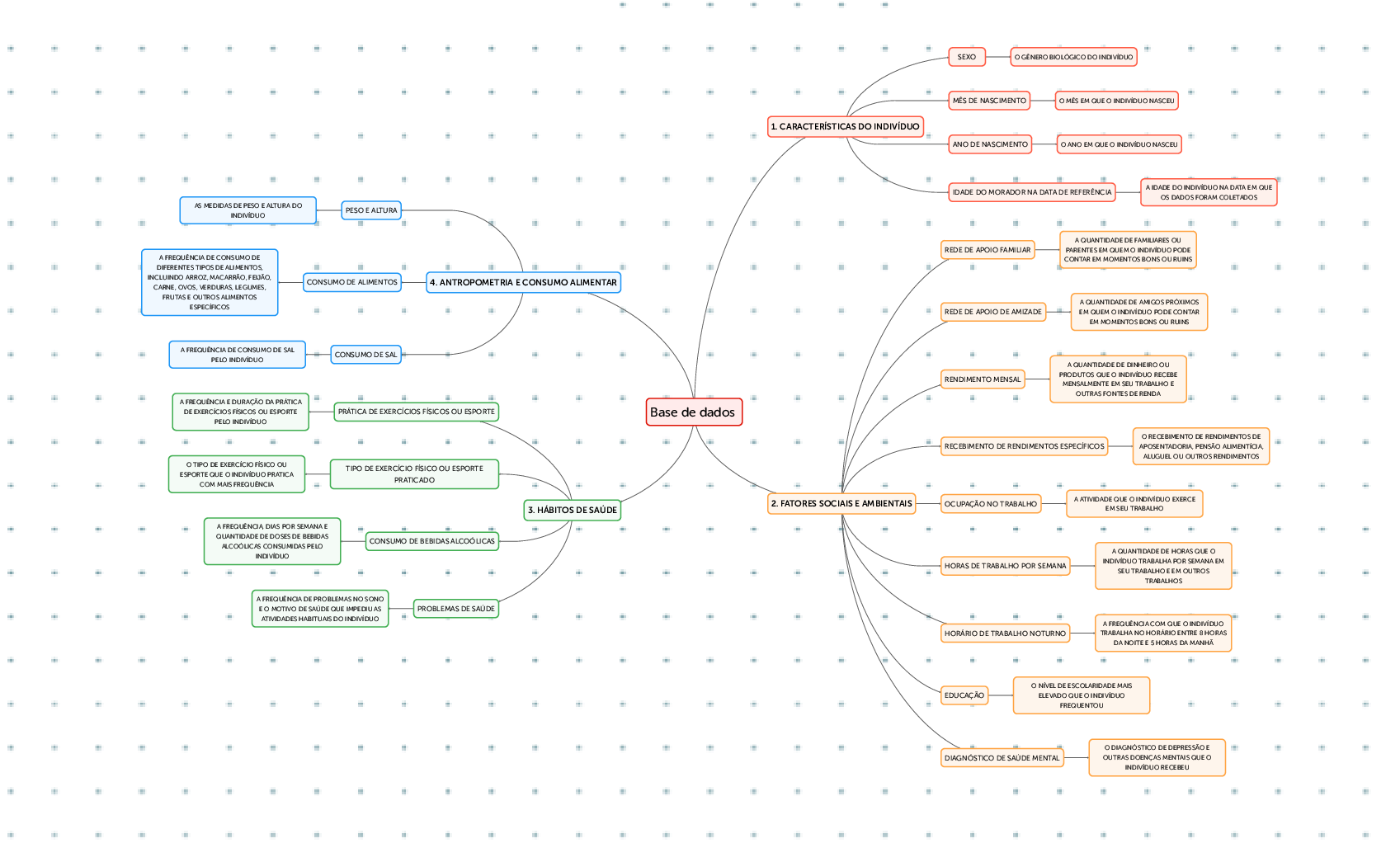


Figura 1 – Descrição dos Atributos da Base de Dados

**3. Etapas de pré-processamento**

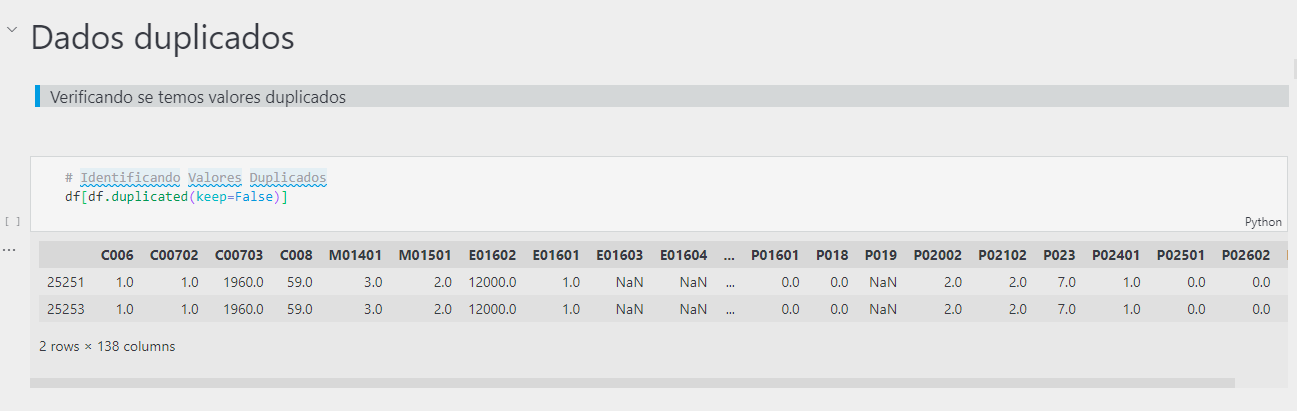
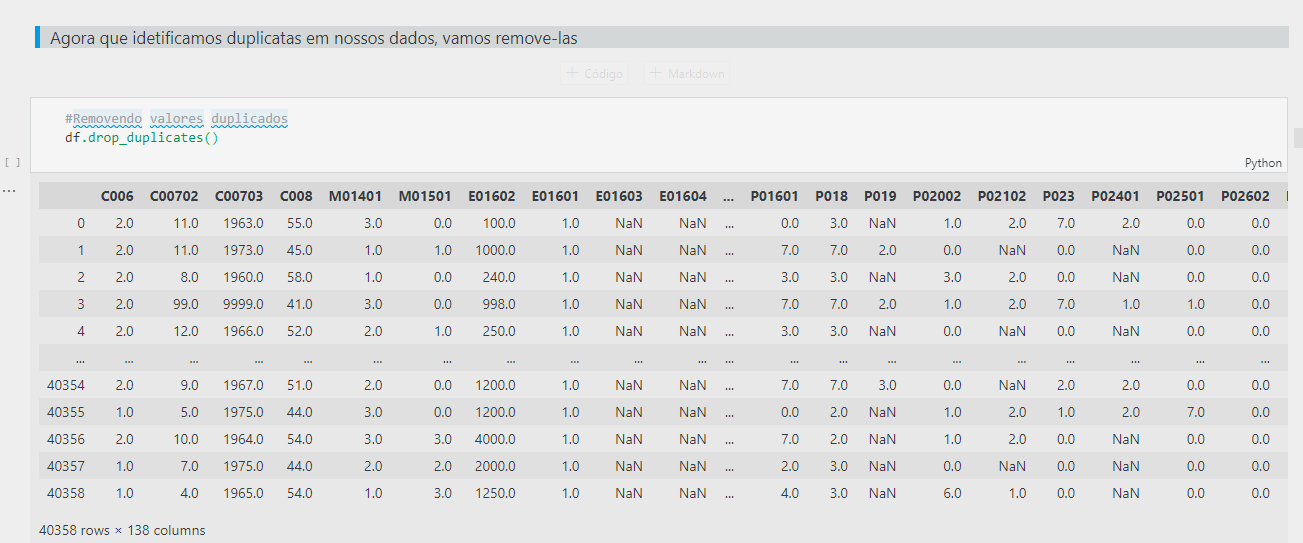
* **Bibliotecas Utilizadas**

**As principais bibliotecas utilizadas nesta etapa são:**

1. **Sidetable: Auxilia na identificação rápida de valores faltantes e frequências de valores categóricos.**
2. **Pandas Profiling: Faz uma análise exploratória automática dos dados, gerando insights sobre problemas nos dados.**
3. **NumPy: Biblioteca fundamental para computação científica em Python.**
4. **Pandas: Biblioteca popular para análise de dados.**
5. **Seaborn / Matplotlib: Utilizadas para visualização de dados.**
6. **Missingno: Facilita a plotagem de dados faltantes.**
7. **IPython Widgets: Permite interatividade.**

**Essas bibliotecas forneceram as ferramentas necessárias para realizar as análises e o pré-processamento dos dados de forma eficiente e eficaz.**

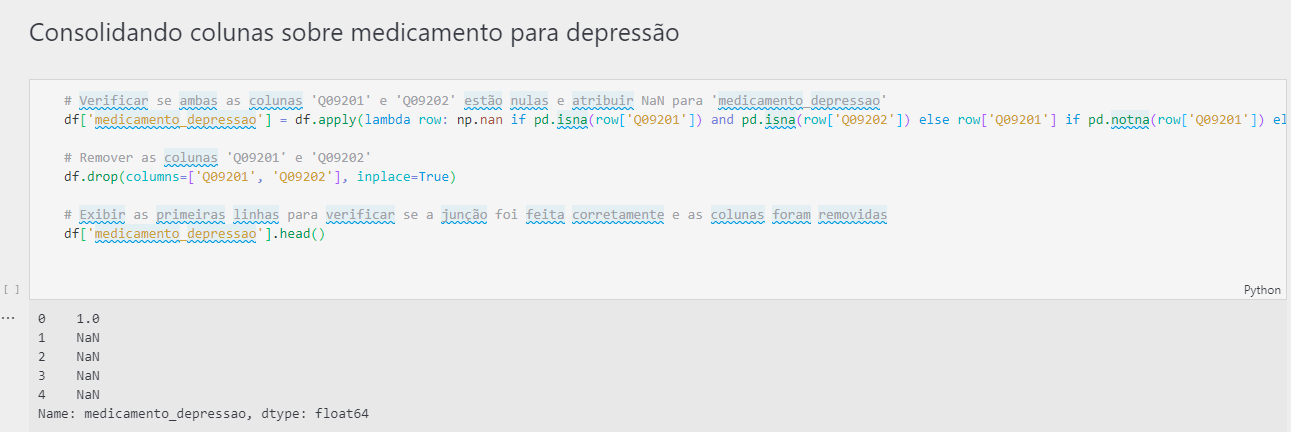
1. **Remoção de Atributos Duplicados:**



1. **Consolidando Colunas** 

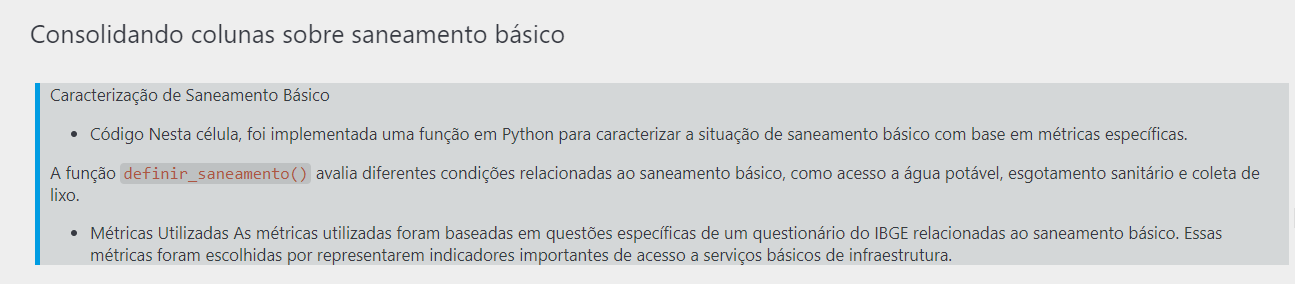




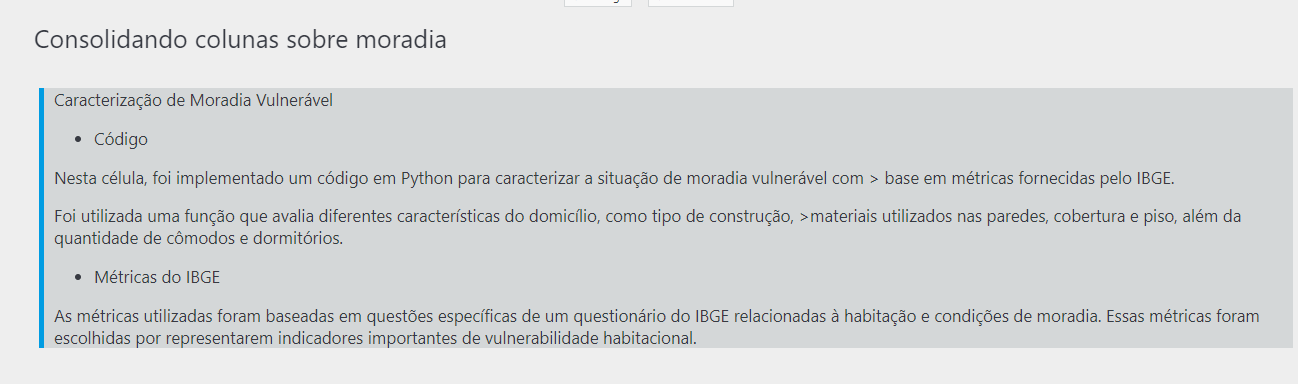


****

****

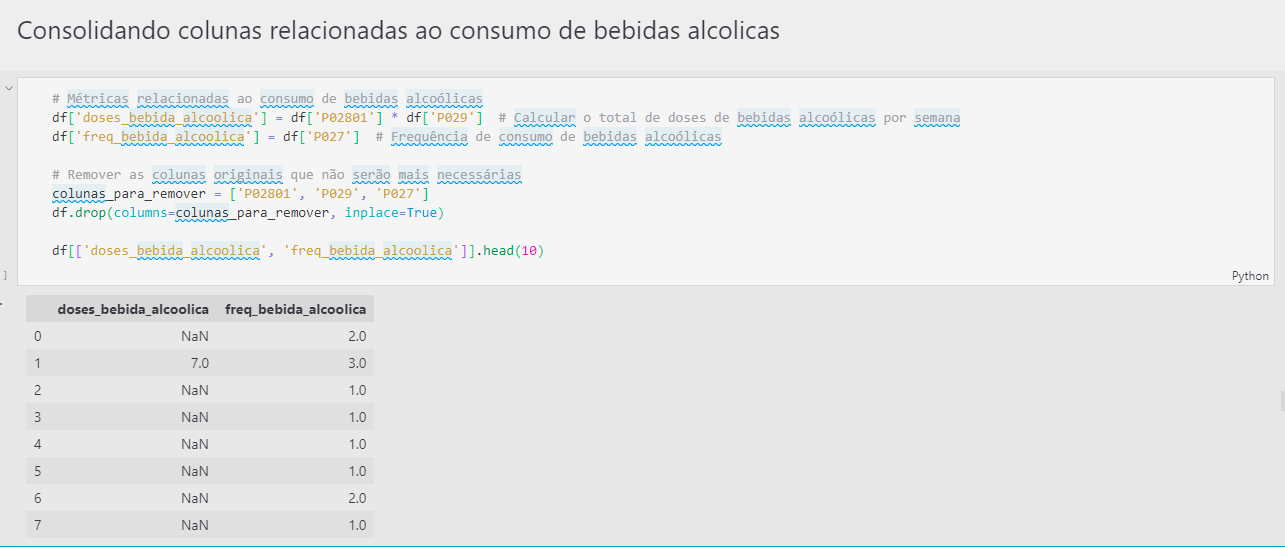
****

****

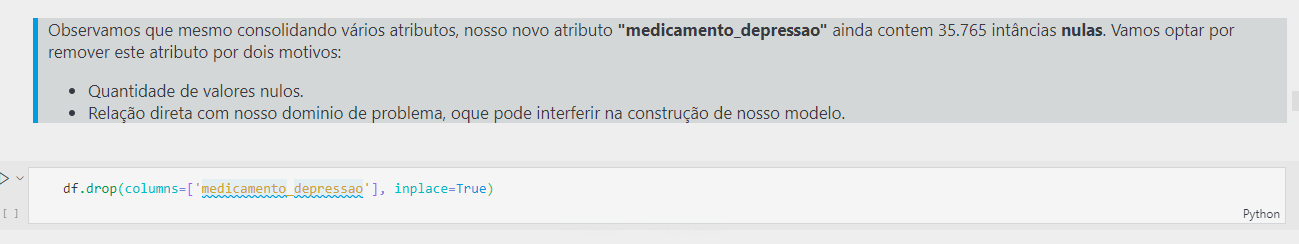
****

****

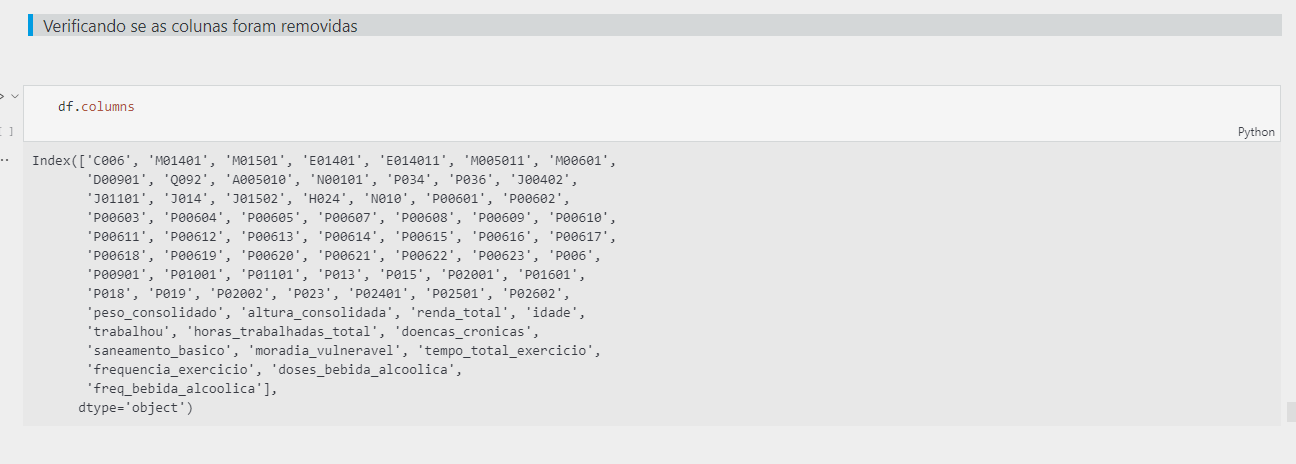
****

****

1. **Removendo Colunas Irrelevantes:**

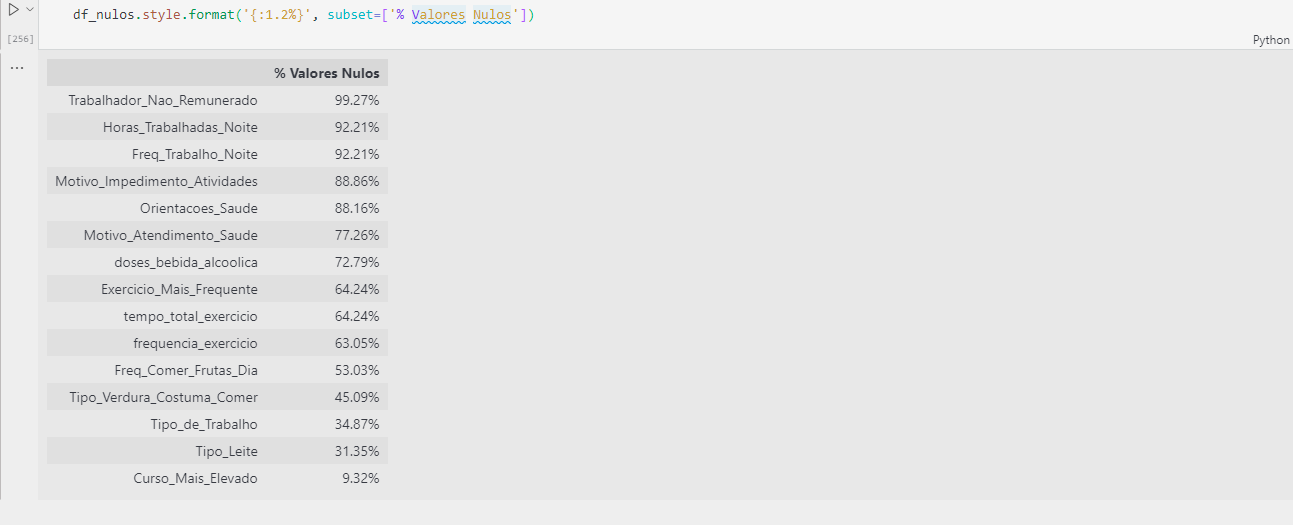
****

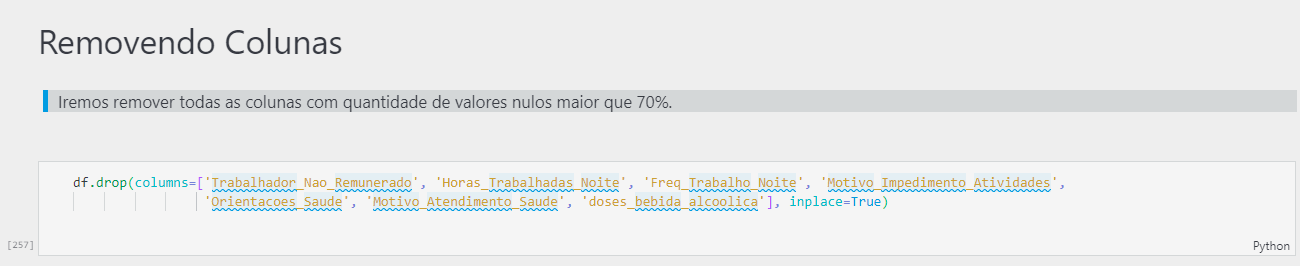
****

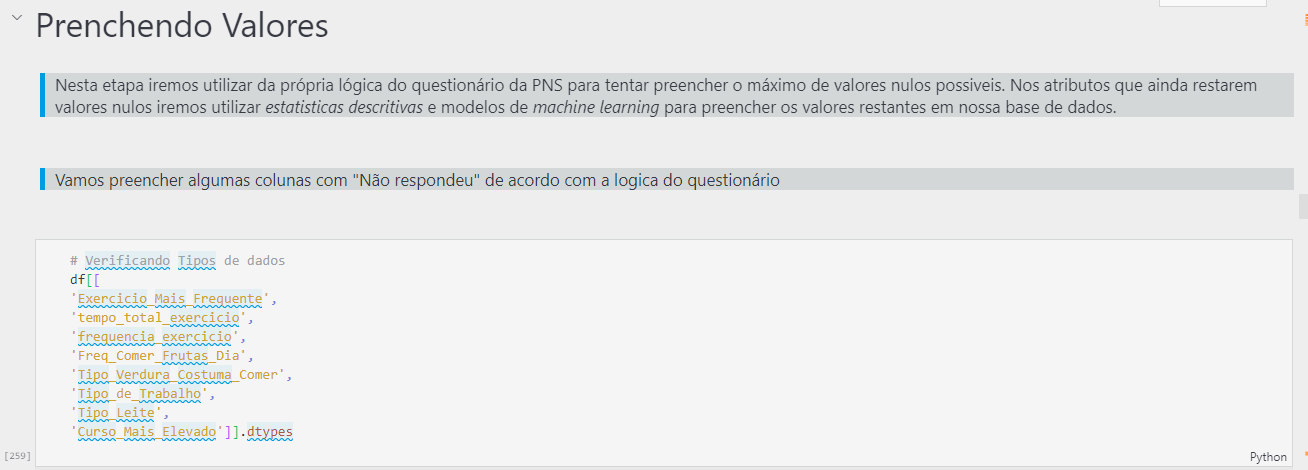


1. **Removendo Valores Nulos:**

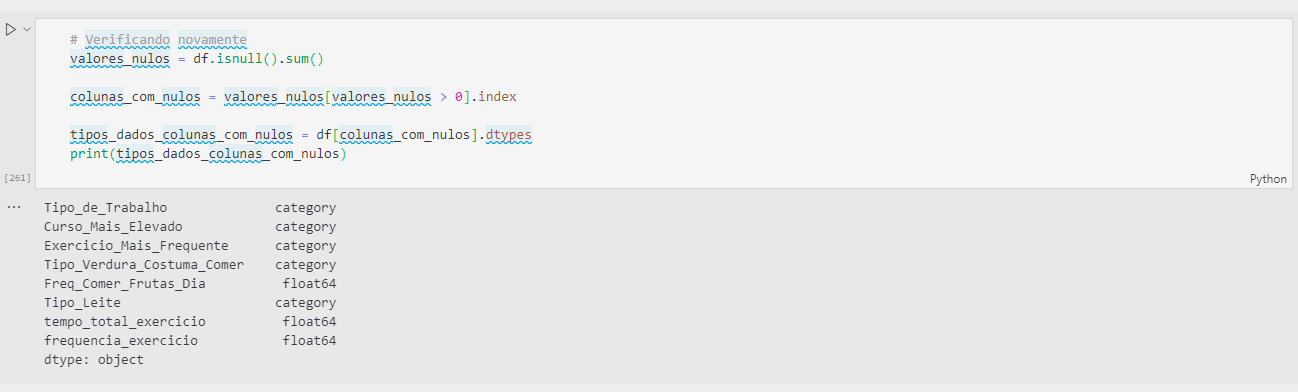
****

****

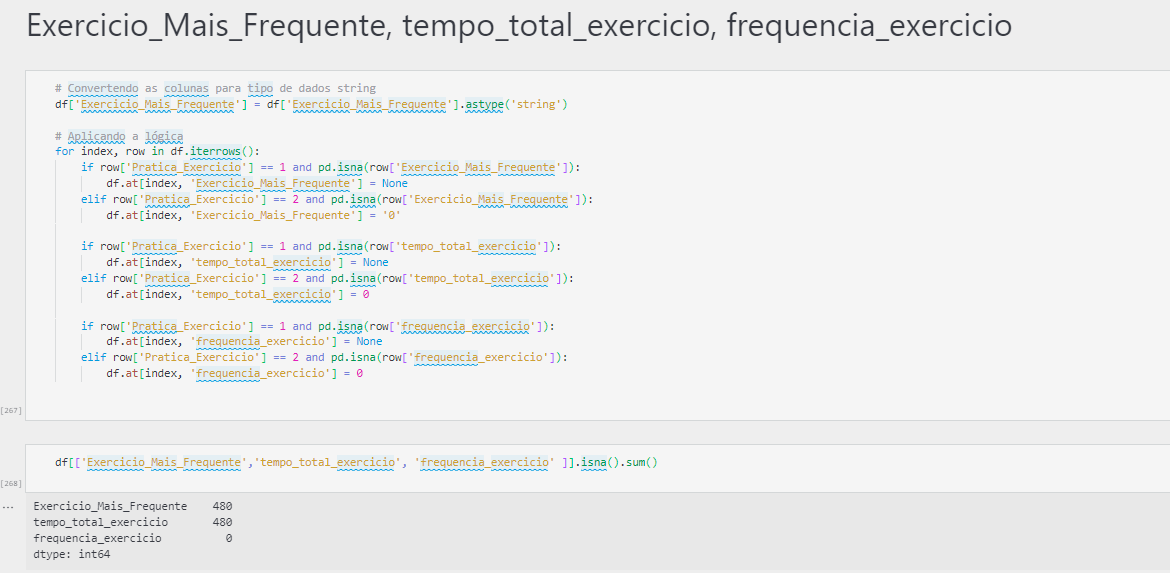
****

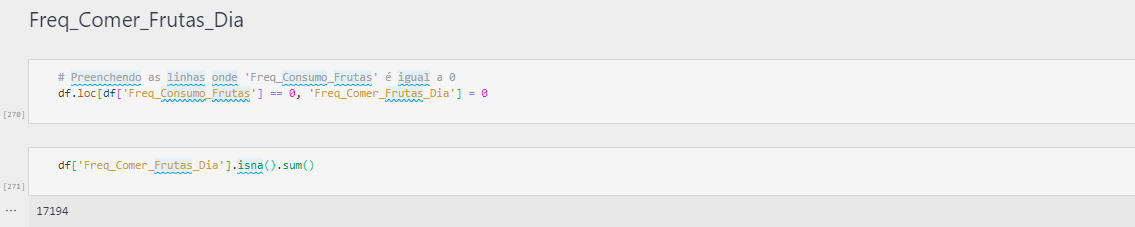
****

****

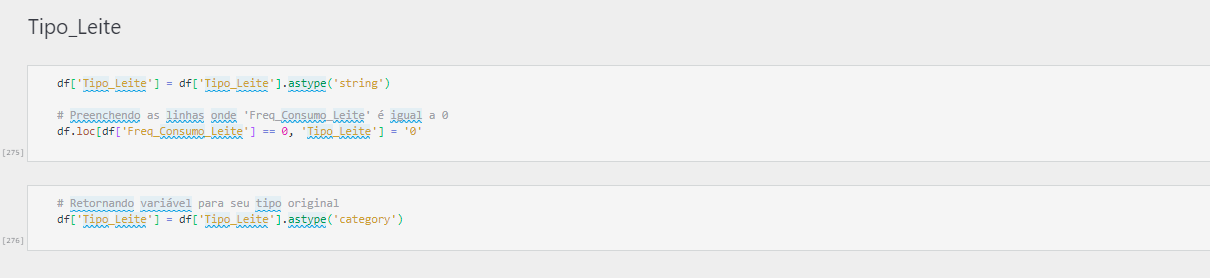
****

****

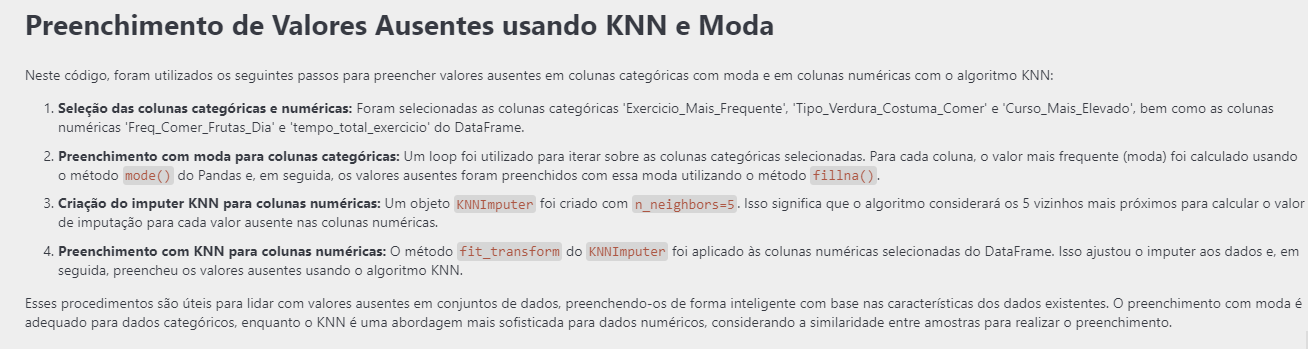
****

****

****

****

****

****

****

****

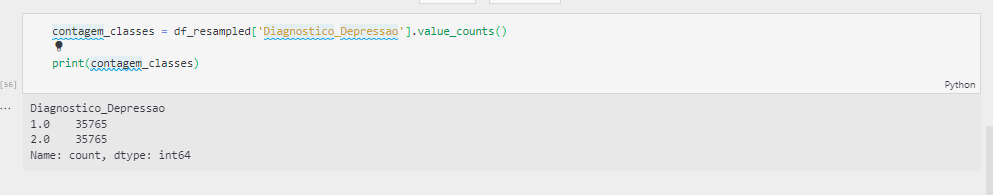
****

A avaliação dos resultados apresentados indica que a acurácia para preenchimento das colunas numéricas e categóricas foi de 100%. Isso sugere que o processo de imputação de dados ausentes utilizando moda para colunas categóricas e o algoritmo KNN para colunas numéricas obteve resultados muito bons.

Com uma acurácia de 100% para todas as variáveis, podemos inferir que o processo de imputação foi bem-sucedido e que os valores ausentes foram preenchidos de forma precisa e eficiente. Isso é importante para garantir a qualidade dos dados e evitar viés nos resultados das análises posteriores.

1. **Balanceamento dos dados**





Para equilibrar os dados, utilizamos o método de oversampling com a técnica RandomOverSampler, implementada pela biblioteca imbalanced-learn (imblearn). Abaixo está uma explicação detalhada do processo:

* Equilíbrio dos Dados usando RandomOverSampler
* Importação da Biblioteca: Utilizamos a biblioteca imbalanced-learn para implementar a técnica de oversampling.
* Criação do Objeto de Oversampling: Criamos um objeto chamado oversampler utilizando a classe RandomOverSampler. Essa classe é responsável por equilibrar os dados aplicando a técnica de oversampling, que consiste em aumentar a quantidade de instâncias da classe minoritária.
* Separação das Features e do Target: Separamos as features (X) e o target (y) do DataFrame original. As features representam as variáveis independentes do modelo, enquanto o target representa a variável dependente que queremos prever ou classificar.
* Aplicação do Oversampling: Aplicamos o método fit\_resample ao conjunto de features (X) e target (y) utilizando o objeto oversampler. Esse método realiza o equilíbrio dos dados, gerando novas amostras para a classe minoritária de forma aleatória até que as classes estejam equilibradas.
* Criação do DataFrame Equilibrado: Após o equilíbrio, combinamos novamente as features e o target em um novo DataFrame chamado df\_resampled. Esse DataFrame agora contém um número igual de amostras para cada classe, garantindo que o conjunto de dados esteja equilibrado.

**Hyperparâmetros Utilizados:**

Não especificamos hyperparâmetros específicos, pois o RandomOverSampler não requer ajustes adicionais além da classe majoritária e minoritária.

**4. Descrição dos métodos utilizados:**

Para abordar o problema de classificação do diagnóstico de depressão, foram utilizados dois algoritmos de aprendizado de máquina: Naive Bayes e Árvore de Decisão.

1. Naive Bayes:

- O algoritmo de Naive Bayes é um método de classificação baseado no teorema de Bayes com a suposição "ingênua" de independência condicional entre os recursos.

- Utilizamos a implementação do algoritmo Gaussiano (`GaussianNB`) do pacote `sklearn`.

- Não foi necessário ajustar hiperparâmetros específicos neste caso, pois o Naive Bayes é conhecido por sua simplicidade e eficiência em muitos problemas de classificação.

2. Árvore de Decisão:

- O algoritmo de Árvore de Decisão é um método de aprendizado supervisionado que aprende uma série de regras de decisão simples com base nos recursos dos dados.

- Utilizamos a implementação do algoritmo de Árvore de Decisão (`DecisionTreeClassifier`) do pacote `sklearn`.

- Neste modelo, o hiperparâmetro `random\_state` foi definido como 42 para garantir a reprodutibilidade dos resultados.

Ambos os modelos foram implementados utilizando a linguagem de programação Python, especificamente com o auxílio das bibliotecas `pandas`, `scikit-learn`, `matplotlib` e `seaborn`. O pré-processamento dos dados, como lidar com valores ausentes e converter variáveis categóricas em numéricas, foi realizado antes do treinamento dos modelos.

**Melhor modelo:**

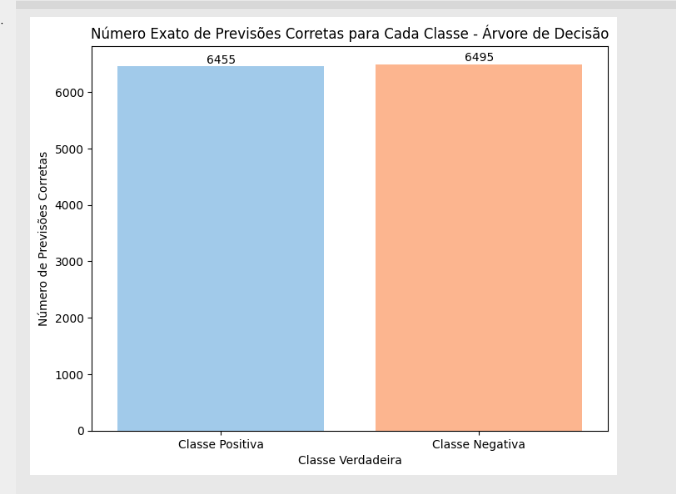
Para determinar qual modelo é o melhor para resolver o problema de classificação do diagnóstico de depressão, devemos considerar não apenas as métricas de desempenho, mas também as necessidades específicas do problema. Após avaliação dos resultados, o modelo de Árvore de Decisão obteve uma acurácia ligeiramente superior (90.52%) em comparação com o Naive Bayes (88.93%). Além disso, o modelo de Árvore de Decisão demonstrou uma precisão, recall e pontuação F1 ligeiramente melhores em comparação com o Naive Bayes. Portanto, com base nessas métricas de desempenho, o modelo de Árvore de Decisão pode ser considerado o melhor para este problema específico.

**Discussão dos resultados:**

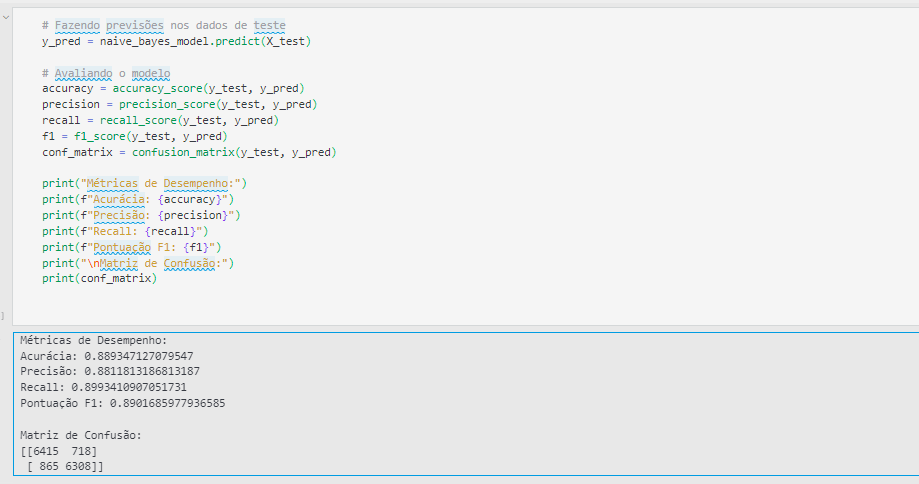
Ambos os modelos, Naive Bayes e Árvore de Decisão, demonstraram desempenho promissor na classificação do diagnóstico de depressão. No entanto, o modelo de Árvore de Decisão mostrou um desempenho ligeiramente superior em termos de métricas de desempenho, como acurácia, precisão, recall e pontuação F1. Isso sugere que a estrutura de decisão mais complexa da Árvore de Decisão pode capturar melhor as relações entre os recursos e a variável alvo.

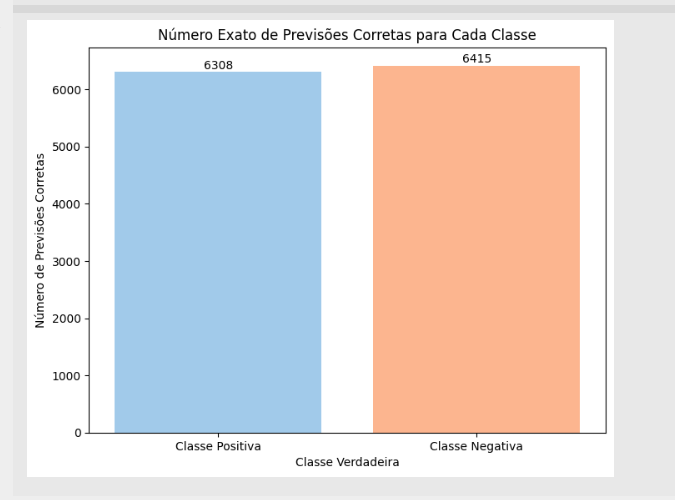
**Avaliação do modelo de Árvore de Decisão:**

****

****

**Avaliação do Modelo Naive Bayes:**

****

****

**5. Código desenvolvido**

https://github.com/Pedro-io/projeto-aprendizado-de-maquina.git

**REFERENCES**

1. Dondena, L. M., Ferretti, E., Maragoudakis, M., Sapino, M., & Errecalde, M. L. (2017). Predicting Depression: a comparative study of machine learning approaches based on language usage. Cuadernos de Neuropsicologia, 11, 42-54.
2. Batista, H. M. C. da, Paim, A. B., Siqueira, B. S., Ebecken, N. F. F., & Dias, A. C. (n.d.). Fatores que podem desencadear depressão: uma aplicação do aprendizado de máquina aos dados da Pesquisa Nacional de Saúde no Brasil.
3. Liu, Y., Pu, C., Xia, S., Deng, D., Wang, X., & Li, M. (2022). Machine learning approaches for diagnosing depression using EEG: A review. Transl Neurosci, 13(1), 224-235.
4. Patel, M. J., & Khalaf, A. (2015). Studying depression using imaging and machine learning methods. NeuroImage: Clinical.
5. Umberson, D., & Montez, J. K. (2010). Social relationships and health: A flashpoint for health policy. Journal of health and social behavior, 51(1\_suppl), S54-S66.
6. Teo, A. R., Choi, H., & Valenstein, M. (2013). Social relationships and depression: Ten-year follow-up from a nationally representative study. PloS one, 8(4), e62396.
7. Whisman, M. A. (2010). The association between depression and marital dissatisfaction. In Handbook of depression (pp. 165-180). Guilford Press.
8. Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. American journal of epidemiology, 157(2), 98-112.
9. Stansfeld, S., & Candy, B. (2006). Psychosocial work environment and mental health—a meta-analytic review. Scandinavian journal of work, environment & health, 32(6), 443-462.
10. Virtanen, M., Stansfeld, S. A., Fuhrer, R., & Ferrie, J. E. (2018). Overtime work as a predictor of major depressive episode: a 5-year follow-up of the Whitehall II study. PLoS One, 13(8), e0202224.
11. Kessler, R. C., Davis, C. G., & Kendler, K. S. (1995). Childhood adversity and adult psychiatric disorder in the US National Comorbidity Survey. Psychological medicine, 25(1), 51-67.
12. Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). Cognitive therapy of depression. Guilford press.
13. Kendler, K. S., Gatz, M., Gardner, C. O., & Pedersen, N. L. (2006). A Swedish national twin study of lifetime major depression. American Journal of Psychiatry, 163(1), 109-114.
14. Smith, K. J., Victor, C., & Bartholomew, J. (2006). Factors associated with the self-reported health status of older people in the United Kingdom. Ageing & society, 26(4), 607-627.
15. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. Archives of general psychiatry, 4(6), 561-571.
16. Kuehner, C. (2017). Gender differences in unipolar depression: an update of epidemiological findings and possible explanations. Acta Psychiatrica Scandinavica, 95(3), 163-174.
17. Fiske, A., Wetherell, J. L., & Gatz, M. (2009). Depression in older adults. Annual Review of Clinical Psychology, 5, 363-389.
18. Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: Review and meta-analysis. American journal of psychiatry, 157(10), 1552-1562.
19. Young, E. A., Korszun, A., & Sex and Gender in Mood Disorders Task Force. (2014). The hypothalamic-pituitary-gonadal axis in mood disorders. Endocrinology and metabolism clinics of North America, 43(3), 683-696.
20. Luppino, F. S., de Wit, L. M., Bouvy, P. F., Stijnen, T., Cuijpers, P., Penninx, B. W., & Zitman, F. G. (2010). Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. Archives of general psychiatry, 67(3), 220-229.
21. Lai, H. M. X., Cleary, M., Sitharthan, T., & Hunt, G. E. (2015). Prevalence of comorbid substance use, anxiety and mood disorders in epidemiological surveys, 1990–2014: A systematic review and meta-analysis. Drug and alcohol dependence, 154, 1-13.
22. Skogen, J. C., Harvey, S. B., Henderson, M., Stordal, E., Mykletun, A., & Øverland, S. (2014). Anxiety and depression among abstainers and low-level alcohol consumers: The Nord-Trøndelag Health Study. Addiction, 109(2), 269-277.