**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina**

**Salatiel Costa Bairros**

**Allan Foppa Fagundes**

**RELATÓRIO DE FELICIDADE MUNDIAL: ESTIMANDO A FELICIDADE A PARTIR DE ÍNDICES SOCIAIS E ECONÔMICOS UTILIZANDO APRENDIZADO DE MÁQUINA**

Belo Horizonte

Abril de 2022

**Salatiel Costa Bairros**

**Allan Foppa Fagundes**

**RELATÓRIO DE FELICIDADE MUNDIAL: ESTIMANDO A FELICIDADE A PARTIR DE ÍNDICES SOCIAIS E ECONÔMICOS UTILIZANDO APRENDIZADO DE MÁQUINA**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina, como requisito parcial à obtenção do título de *Especialista*.

Belo Horizonte

Junho de 2022

**SUMÁRIO**

[1. Introdução 4](#_Toc105261013)

[2. Descrição do Problema e da Solução Proposta 4](#_Toc105261014)

[3. Canvas Analítico 6](#_Toc105261015)

[4. Coleta de Dados 8](#_Toc105261016)

[5. Processamento/Tratamento de Dados 11](#_Toc105261017)

[6. Análise e Exploração dos Dados 16](#_Toc105261018)

[7. Preparação dos Dados para os Modelos de Aprendizado de Máquina 17](#_Toc105261019)

[8. Links 17](#_Toc105261020)

[9. Referências 18](#_Toc105261021)

# 1. Introdução

O campo de Inteligência Artificial e Aprendizado de Máquina tem sido cada vez mais utilizado por organizações dos setores público e privado identificando fraudes no uso do dinheiro público e comportamento de clientes para otimizar vendas. Cada vez mais interdisciplinar, a área é composta por matemáticos, estatísticos, cientistas da computação, linguistas, biólogos, cientistas sociais e muitos outros. Somado a isso, o avanço do poder computacional e a criação de ferramentas que permitem a análise e modelagem dos dados para o Aprendizado de Máquina contribuem para o grande sucesso dessa área da Ciência da Computação.

Em seu intenso e amplo uso, a Inteligência Artificial enfrenta hoje, entre outros, dois grandes desafios: (1) poder e custo computacional de execução em modelos complexos e (2) qualidade e quantidade dos dados necessários para o treinamento de bons modelos. O primeiro tem sido enfrentado com o uso de Unidades de Processamento Gráfico (GPUs), Computação em Nuvem e pesquisas com Computação Quântica. Já o segundo tem encontrado soluções através da colaboração com as áreas de conhecimento especialistas nos dados desejados. Essa coleta pode ser feita de modo manual, com pesquisa de campo, coleta de dados e entrevistas, ou de forma automatizada, lendo dados da *Internet*, extraindo dados de textos, áudios, vídeos, entre outros.

Beneficiando-se dessa interdisciplinaridade, o presente trabalho utilizará o Relatório de Felicidade Mundial, uma pesquisa realizada por uma iniciativa das Nações Unidas conhecida como *Sustainable Development Solutions Network*. O propósito é identificar a relação entre os índices apresentados e o *score* de felicidade dos países.

# 2. Descrição do Problema e da Solução Proposta

Dentre todos os aspectos da vida valorizados no mundo moderno, é possível afirmar que a felicidade está entre os principais. Com o propósito de facilitar a análise da felicidade social como indicador de desenvolvimento, a Organização das Nações Unidas, por meio da Rede de Soluções para o Desenvolvimento Sustentável, criou o Relatório de Felicidade Mundial, publicado anualmente baseado nos dados do *Gallup World Poll.*

A criação do relatório foi inspirada no Butão que desde 1972 utiliza um indicador de felicidade para complementar as métricas tradicionais para acompanhar o desenvolvimento do país: a Felicidade Interna Bruta, composta por 9 categorias: bem-estar psicológico, saúde, uso do tempo, vitalidade comunitária, educação, cultura, meio ambiente, governança e padrão de vida.

Em 2011 a ONU recomendou aos países participantes que também utilizassem indicadores de felicidade para acompanharem o seu desenvolvimento e em 2012 o relatório foi lançado, sendo anualmente atualizado e com os dados disponibilizados de forma pública.

Diante disso, o presente trabalho tem por objetivo principal validar a possibilidade de construir um modelo de Aprendizado de Máquina que seja capaz de, dado os índices utilizados no relatório, estimar o valor do índice de felicidade em um determinado país. Além disso, como objetivos secundários, analisar a existência de alterações nos valores causados pela pandemia (2020 e 2021) e identificar se é possível, através de modelos de classificação, determinar qual a região no mundo em que um país se encontra através da relação entre as métricas da pesquisa e o índice de felicidade do país.

Serão utilizados algoritmos de aprendizado de máquina para regressão, classificação e clusterização dos dados. Os algoritmos de regressão, como Regressão Linear, Regressão Bayesiana, KNN, Árvore de Decisão e Redes Neurais, serão utilizados para o objetivo principal: obter um valor do índice de felicidade dado as características de determinado país. Os modelos serão individualmente analisados e o melhor modelo será selecionado para o *pipeline* de publicação em produção. Enquanto os algoritmos de classificação, como Árvore de Decisão, Redes Neurais e SVM serão utilizados para o objetivo secundário de classificar um país em uma região através das informações fornecidas. Por fim, os algoritmos de clusterização, como *K-Means, DBScan* e *MeanShift* serão utilizados na etapa de análise de dados, buscando melhor encontrar as relações entre os dados e a melhor forma de realizar o pré-processamento do modelo.

A linguagem escolhida para desenvolver os modelos de Aprendizado de Máquina juntamente com as análises e transformações de dados necessárias é Python, dado a sua grande popularidade no uso para aplicações do tipo e o grande número de bibliotecas de código aberto que implementam os algoritmos propostos. Algumas das bibliotecas utilizadas são: *sklearn*, *keras*, *numpy* e *pandas*.

Por fim, os dados, as análises e o código fonte serão disponibilizadas na plataforma GitHub, tendo os resultados disponibilizados via REST API.

# 3. Canvas Analítico

Com o objetivo de mapear de forma clara e objetiva o propósito do projeto, fez-se a escolha de utilizar o *Software Analytics Canvas* criado pelo Analista de Desenvolvimento de Software Markus Harrer, adaptando o modelo para as necessidades do projeto. Serão utilizadas as seguintes etapas do Canvas: questões, fontes de dados, hipóteses (como substituto de heurísticas), validação e implementação. Abaixo a definição de cada uma dessas etapas com seu respectivo papel para realização do projeto.

**3.1 Questões**

A etapa de questões corresponde às perguntas que serão feitas sobre os dados, buscando obter informações relevantes, assim como identificar e resolver problemas apontados por eles. As questões que o presente trabalho estipulou são baseadas nos objetivos principais e secundários apresentados. Sendo elas:

1. É possível estimar a felicidade média de um país através de métricas quantificáveis de desenvolvimento humano?
2. É possível identificar a região no mundo em que um país se encontra através da relação entre as métricas e o índice de felicidade obtida?
3. A pandemia causou algum impacto nos índices obtidos de felicidade mundial?

**3.2 Fontes de dados**

Conforme introduzido na seção 2 (Descrição do Problema e da Solução Proposta) e detalhado abaixo na seção 4 (Coleta de Dados), serão utilizados os dados disponibilizados no Relatório de Felicidade Mundial de 2021, realizado pela iniciativa *Sustainable Development Solutions Network* da ONU com os dados da *Gallup World Poll.*

**3.3 Hipóteses**

A terceira etapa, originalmente chamada de “Heurísticas” foi adaptada ao problema do projeto, baseado na própria definição original, que a definia como as suposições realizadas para facilitar a resposta das questões.

A hipótese central que motiva a análise dos dados e sua respectiva modelagem em algoritmos de Aprendizado de Máquina é de que métricas quantificáveis de qualidade de vida têm influência na percepção de felicidade das pessoas, porém pode ser relativa à própria qualidade de vida descrita nas métricas que, dado um longo período, leva as pessoas a se acostumarem com elas e elevarem o seu padrão de vida ideal, alterando a própria percepção de felicidade. Ou seja, o índice de felicidade não é completamente explicado puramente pelas métricas, ainda que possua uma relação com elas.

**3.4 Validação**

A penúltima etapa aplicada se refere, segundo Harrer, à forma com que os resultados serão disponibilizados e apresentados de maneira fácil de compreender. Baseado nisso, a proposta deste trabalho é apresentar os resultados da análise e modelagem em uma página web e API pública.

**3.5 Implementação**

Por fim, a etapa de implementação responde à pergunta de como o projeto será implementado. Ela será dividida em quatro partes. São elas:

1. Preparação dos dados para análise e modelagem
2. Análise gráfica e estatística dos dados coletados
3. Utilização de algoritmos de aprendizagem de máquina buscando responder as perguntas iniciais.
4. Disponibilização dos resultados no formato informado na etapa de validação.

É necessário destacar, contudo, que as partes não são concluídas necessariamente de forma sequencial, visto que os resultados de uma parte podem afetar o desenvolvimento da anterior até que o projeto seja inteiramente concluído.

# 4. Coleta de Dados

Conforme mencionado anteriormente, os dados foram obtidos por meio do Relatório de Felicidade Mundial, disponível publicamente e coletados pela empresa *Gallup* através do *Gallup World Poll*, tendo como propósito a criação de um índice de desenvolvimento que não fosse exclusivamente atrelado a métricas financeiras, como o PIB, buscando registrar a qualidade de vida das pessoas por informações mais subjetivas, como a felicidade. Contendo dados anuais entre 2008 e 2021 coletados na maior parte dos países pertencentes à ONU, os dados se mostram muito relevantes para acompanhar e comparar a percepção mundial de felicidade ao longo do tempo e entre países e regiões do mundo.

As informações são divididas em dois arquivos no formato *csv* com os dados: um contendo os dados até 2020 e outro apenas com os dados de 2021. A estrutura de ambos os arquivos é bastante similar, exceto pelos atributos “*Positive affect”* e “*Negative affect”*, ausentes nos dados de 2021 e “*Regional indicator”*, presente apenas nos dados de 2021. Atributos relativos a métricas estatísticas, como “*Standard error of ladder score”,* foram ignorados por divergirem entre os arquivos, não permitindo o uso para análise e modelagem.

Nas tabelas abaixo são listados os atributos dos dois *datasets* utilizados. Como a maioria dos atributos está presente nos dois arquivos, a primeira tabela conterá todos os atributos utilizados de 2021 e a segunda, correspondente aos dados até 2020, apenas com o que for ausente nos dados de 2021.

|  |  |  |
| --- | --- | --- |
| **Nome do dataset:** World Happiness Score 2021  **Descrição:** Registros das informações do índice de felicidade do ano de 2021.  **Link:** https://happiness-report.s3.amazonaws.com/2021/DataForFigure2.1WHR2021C2.xls | | |
| **Nome do Atributo** | **Descrição** | **Tipo** |
| Country name | Nome do país onde os dados foram obtidos. | Texto |
| Regional indicator | Região do país no mundo. Apenas nos dados de 2021. | Texto |
| Ladder score | Média da percepção dos entrevistados sobre sua vida em uma escala de 0 a 10. | Decimal |
| Logged GDP per capita | Indicador que mostra a paridade de compra per capita, tendo o dólar como base. | Decimal |
| Social support | Cada entrevistado responde se ele possui pessoas que poderia contar em um eventual momento de necessidade. O valor 1 é sim e 0 é não. O resultado é a média dessas respostas. | Decimal |
| Healthy life expectancy (HLE) | O número médio de anos esperados de vida com plena saúde do país. | Decimal |
| Freedom to make life choices | O entrevistado responde se percebe a liberdade de escolher o que é melhor para si, sendo 1 para sim e 0 para não. O valor final é a média das respostas. | Decimal |
| Generosity | O entrevistado responde se doou para caridade no último mês. O valor 1 é sim e 0 é não. O resultado é a média dessas respostas. | Decimal |
| Perceptions of corruption | O entrevistado responde se percebe corrupção generalizada no setor privado e no setor público. O valor 1 é sim e 0 é não. O resultado é a média das respostas | Decimal |

|  |  |  |
| --- | --- | --- |
| **Nome do dataset:** World Happiness Score até 2020  **Descrição:** Registros das informações do índice de felicidade obtidos até 2020, disponibilizados no relatório de 2021.  **Link:** https://happiness-report.s3.amazonaws.com/2021/DataPanelWHR2021C2.xls | | |
| **Nome do Atributo** | **Descrição** | **Tipo** |
| Positive affect | Cada entrevistado responde se, durante as últimas 48 horas, sentiu felicidade, apreciação ou sorriu bastante. É feita uma média desses três fatores para o entrevistado e o valor final é a média de todos os entrevistados do país. | Decimal |
| Negative affect | Semelhante ao *Positive affect,* mas com sentimentos negativos: preocupação, tristeza e raiva. | Decimal |
| year | Ano da coleta dos dados | Inteiro |

Além dos dois *datasets* principais com os dados foi também utilizado um dataset para categorizar e relacionar corretamente os países às suas respectivas regiões do mundo. Abaixo o *dataset* com as colunas utilizadas no tratamento:

|  |  |  |
| --- | --- | --- |
| **Nome do dataset:** Countries of the World  **Descrição:** Informações gerais sobre os países do mundo.  **Link:** https://www.kaggle.com/datasets/fernandol/countries-of-the-world?select=countries+of+the+world.csv | | |
| **Nome do Atributo** | **Descrição** | **Tipo** |
| Country | Nome do país | Texto |
| Region | Região do país no mundo | Texto |

# 

# 5. Processamento/Tratamento de Dados

O processamento inicial dos dados realizado se deu em seis etapas, cujo objetivo final era obter um *dataset* único e consistente dos dados, ou seja, sem dados faltantes e com os dados corretamente preenchidos.

As principais bibliotecas utilizadas para manipulação dos dados realizada nas etapas de processamento foram *pandas* e *numpy*, ambas para a linguagem Python e cada etapa se propõe a resolver um problema encontrado nos *datasets*, sendo executadas sequencialmente utilizando uma adaptação do *Commands Design Pattern,* conforme imagem abaixo:

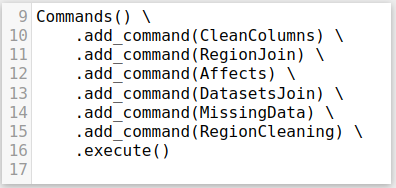


Figura 1 - world-happiness-report/src/data\_preparation\_commands.py

A primeira etapa de preparação e limpeza foi composta pela padronização dos nomes dos atributos e exclusão de atributos que não serão necessários, visto que não são comuns entre os dados de 2021 e os dados históricos.

A segunda etapa teve como objetivo preencher as informações sobre a região do mundo em que cada país se encontra, pois, esta informação está presente apenas nos dados de 2021. Para resolver isso foi realizado o preenchimento dos dados históricos baseados na região informada nos dados do ano de 2021. No entanto, nem todos os países presentes nos dados históricos estão presentes no relatório de 2021 e foram utilizadas bases de dados do Kaggle fornecidas pela própria *Sustainable Development Solutions Network* dos anos de 2015 e 2016. O resultado desta etapa ainda não possui todas as regiões corretamente atribuídas, mas isso será resolvido na última etapa.

A terceira etapa visa lidar com duas colunas ausentes nos dados de 2021 presentes nos dados históricos: “positive affects” e “negative affects”. Ambos os atributos são responsáveis por quantificar o estado sentimental das pessoas nos respectivos países. Para isso é perguntado ao entrevistado a presença de afetos positivos como alegria, apreciação e sorrisos, assim como, sentimentos negativos como preocupação, tristeza e raiva. Devido às questões recentes no cenário mundial nos anos de 2020 e 2021, especialmente a pandemia de COVID-19, optou-se por apenas repetir os valores do ano anterior (2020) para os dados de 2021, ao invés de realizar algum tratamento como a média dos últimos anos, visto que poderia diluir o impacto da pandemia nos resultados. Para os países ausentes no relatório de 2020 os valores foram preenchidos com o ano mais recente cuja informação foi obtida.

A quarta etapa realiza a combinação dos dados históricos com os dados de 2021 em um único *dataset*, dado que ambos estão agora com os mesmos atributos permitindo que as etapas seguintes façam tratamentos mais inteligentes nos dados.

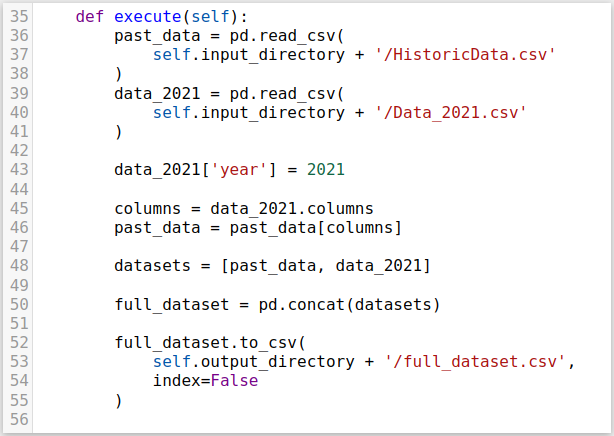


Figura 2 - world-happiness-report/src/data\_preparation/dataset\_join.py

A quinta etapa tem como propósito preencher os dados faltantes nos *datasets* com o menor impacto possível nas relações entre os atributos. A distribuição dos dados faltantes no *dataset* se dá conforme a tabela abaixo:

|  |  |  |
| --- | --- | --- |
| **atributo** | **Nro. faltantes** | **%** |
| corruption | 110 | 5.24 |
| generosity | 89 | 4.24 |
| hle | 55 | 2.62 |
| gdp | 36 | 1.72 |
| freedom | 32 | 1.53 |
| positive\_affect | 22 | 1.05 |
| negative\_affect | 16 | 0.76 |
| social\_support | 13 | 0.62 |

Devido à baixa quantidade de dados faltantes, o algoritmo escolhido para preencher os dados foi:

1. Agrupa-se os dados faltantes de cada atributo por país
2. Todos os países que, no total dos dados obtidos para aquele atributo ao longo do tempo, têm mais de 50% dos dados preenchidos, os faltantes são completados com a média dos anos mais próximos (anterior e posterior).
3. Para os países com 50% ou menos dos dados históricos preenchidos para o atributo atribui-se o menor valor entre a média e a mediana geral deste valor nos dados.

O preenchimento dos dados faltantes seguindo esse algoritmo teve um impacto mínimo nas relações entre os dados, conforme mostra abaixo o mapa de calor das alterações de correlação entre as variáveis com o tratamento dos dados:

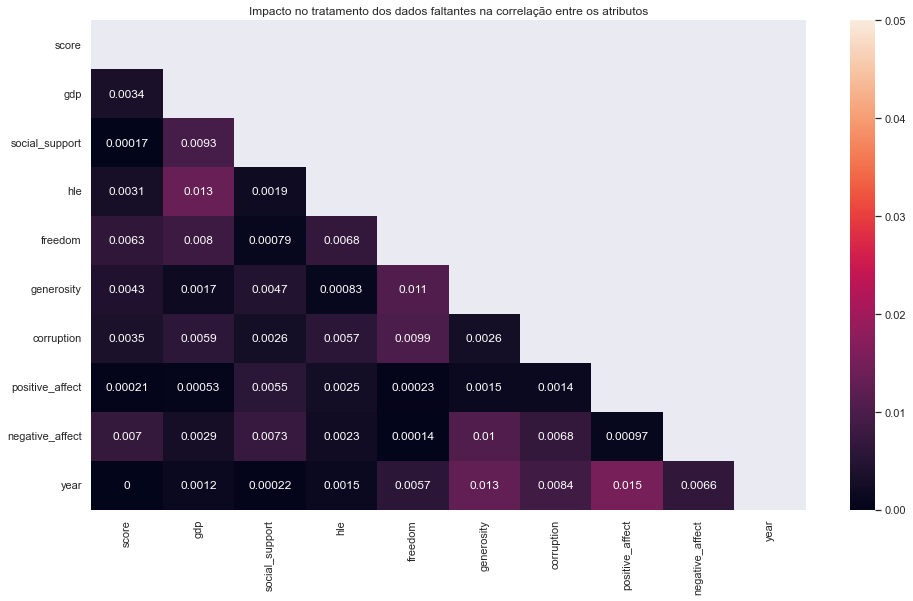


Figura 3 - Mapa de calor baseado no valor absoluto de alteração das correlações entre as variáveis após o tratamento dos dados faltantes.

Mais informações sobre o tratamento dos dados faltantes, análises estatísticas e diferentes abordagens que foram experimentadas para resolver o problema podem ser encontradas em *analysis/MissingData.ipynb* dentro do código-fonte do projeto hospedado no GitHub informado na seção 6.

A sexta e última etapa do processamento inicial dos dados é a categorização correta de cada país em sua respectiva região do mundo utilizando o *dataset* externo informado na seção 4. Para preencher o maior número possível de países com sua região o *dataset* externo passou por alguns tratamentos, conforme código abaixo:



Figura 4 - world-happiness-report/src/data\_preparation/region\_cleaning.py

Por fim, aplicar estes dados no *dataset* principal foram realizados os seguintes passos:

1. Todos os países preenchidos com uma região diferente do *dataset* externo cujo nome foi encontrado no mesmo passam a ter a região encontrada.
2. Países cujo nome não exista com a mesma grafia nos dois *datasets* são encontrados através das suas palavras iniciais ou de conterem alguma palavra presente no *dataset* externo.
3. Regiões específicas que não foram encontradas nos passos acima receberam um mapeamento manual conforme dicionário na imagem abaixo:

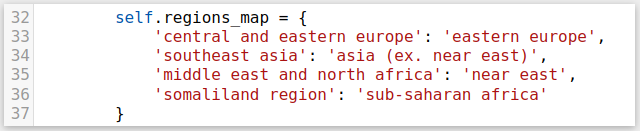


Figura 5 - world-happiness-report/src/data\_preparation/region\_cleaning.py

A implementação desses passos para cada país segue como abaixo:

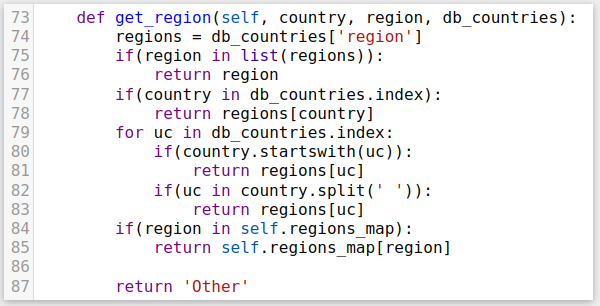


Figura 6 - world-happiness-report/src/data\_preparation/region\_cleaning.py

# 6. Análise e Exploração dos Dados

Nessa etapa você começará a explorar seus dados de uma forma mais analítica, tentando elaborar ideias, levantar hipóteses e começando a identificar padrões em seus dados. Talvez você sinta a necessidade de voltar em passos anteriores, obter mais dados e tratá-los para conseguir responder ao problema proposto. Use e abuse de ferramentas estatísticas consistentes como testes de hipóteses, intervalos de confiança. Plote gráficos que te ajudem a obter insights interessantes: desde os mais simples até gráficos mais sofisticados como boxplots, mapas de calor, etc. Aqui o uso do Python e/ou R e suas poderosas bibliotecas gráficas (Matplotlib, Seaborn, ggPlot2, etc). Apresente trechos de código com as devidas justificativas.

# 7. Preparação dos Dados para os Modelos de Aprendizado de Máquina

Nesta etapa você deve descrever os tratamentos realizados especificamente para os modelos de Aprendizado de Máquina escolhidos, como por exemplo a criação de atributos, o balanceamento da base de dados (*undersampling* ou *oversampling*), divisão da base em treino, validação e teste, entre outros.

# 8. Links

Os arquivos, códigos e artefatos trabalhados no projeto podem ser encontrados em <https://github.com/SalatielBairros/world-happiness-report>.

# 9. Referências

World Happiness Report 2012, acessado 29 março 2022, <<https://worldhappiness.report/>>

Gallup World Poll 2022, acessado 29 Março 2022, <<https://www.gallup.com/178667/gallup-world-poll-work.aspx>>

“Relatório Mundial da Felicidade” Wikipedia, última alteração Março 22, 2022, <<https://pt.wikipedia.org/wiki/Relat%C3%B3rio_Mundial_da_Felicidade>>

ARAUJO, Felipe. Felicidade Interna Bruta (FIB). Março 2022. disponível em: <<https://www.infoescola.com/sociedade/felicidade-interna-bruta-fib/>>

Department of Philosophy, University of Stellenbosch, Stellenbosch, South Africa. 2014, On public happiness, PDF, acessado em 29 Março 2022, <<https://philarchive.org/archive/ROOOPH>>

HARRER, Markus. Feststelltaste. 18 Agosto 2020, Software Analytics Canvas, PDF, acessado em 29 Março 2022, <<https://www.feststelltaste.de/software-analytics-canvas/>>