**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 23](#_Toc105261019)

[8. Links 24](#_Toc105261020)

[9. Referências 25](#_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 |

Tabela 1 – Atributos do dataset com os dados de 2021

|  |  |  |
| --- | --- | --- |
| **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 |

Tabela 2 – Atributos do dataset com os dados até 2020

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 |

Tabela 3 – Atributos do dataset de referência para os países do mundo

# 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:

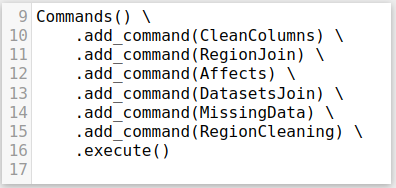


Figura 1 - world-happiness-report/src/data\_preparation\_commands.py

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.

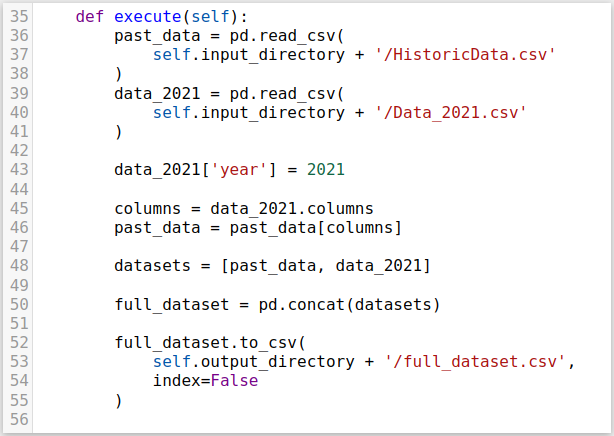


Figura 2 - world-happiness-report/src/data\_preparation/dataset\_join.py

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 |

Tabela 4 – Dados faltantes por atributo

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 (sucessor e antecessor).
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:

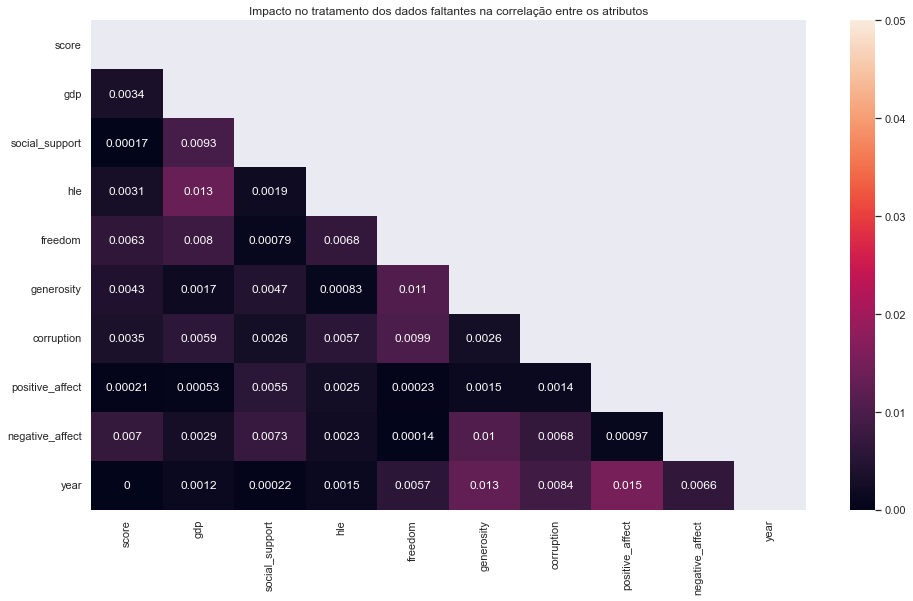


Figura 3 - Mapa de calor baseado no valor absoluto de alteração das correlações entre as variáveis após o tratamento dos dados faltantes.

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:

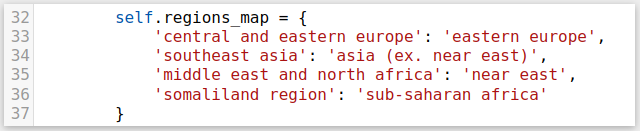


Figura 5 - world-happiness-report/src/data\_preparation/region\_cleaning.py

A implementação desses passos para cada país segue como abaixo:

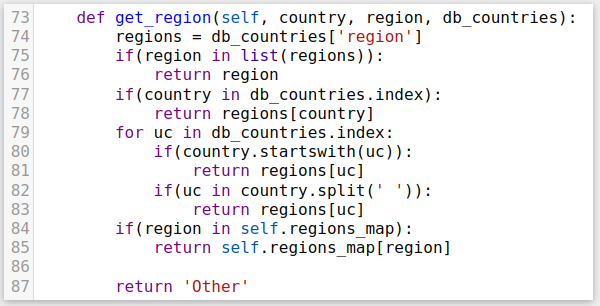


Figura 6 - world-happiness-report/src/data\_preparation/region\_cleaning.py

# 6. Análise e Exploração dos Dados

A análise dos dados teve como principal objetivo identificar possíveis vieses, desbalanceamentos dos dados e indicações da possibilidade de as hipóteses iniciais serem verdadeiras ou não. Para tal, cada atributo foi individualmente analisado, assim como as relações entre os atributos e os resultados foram agrupados em diversos tópicos de análise, considerando a composição dos dados por dois atributos qualitativos categóricos nominais (*country* e *region*), um atributo qualitativo ordinal (*year*) e o restante quantitativos numéricos.

**6.1 Presença dos países e regiões**

Dos 195 países existentes no mundo, 166 estão presentes na pesquisa, ou seja, participaram pelo menos em um dos anos da realização dela. No entanto, a presença ao longo do tempo não foi constante, tendo apenas aproximadamente 29% dos países presentes todos os anos registrados nos dados.

Gráfico

Descrição gerada automaticamente Figura 7 - Presença total dos países na pesquisa.

O gráfico acima destaca a presença de países nas pesquisas realizadas. É notável que a maior parte dos países presentes participaram por mais de 13 anos. Contudo, aproximadamente 30% dos países participaram menos do que isso e tal informação deve ser considerada no momento da construção dos modelos, visto que vieses podem ser gerados por isso.

Uma análise semelhante pode ser realizada para as regiões do mundo e isso acabou por destacar uma informação importante a respeito dos dados: como o número de países é diferente nas regiões do mundo, algumas ficam consideravelmente mais presentes que outras mesmo quando os seus países não possuem individualmente uma presença constante anual nos dados da pesquisa, conforme mostra o gráfico abaixo.

Gráfico, Gráfico de barras

Descrição gerada automaticamente Figura 8 – Presença das regiões do mundo na pesquisa. Note que regiões como a América do Norte possuem baixa presença devido a possuírem poucos países.

A discrepância apontada pela figura acima desafia uma das hipóteses secundárias: identificar a região do mundo através da relação entre o *score* e os índices utilizados. Há uma chance considerável de regiões com menos presença não serem classificadas corretamente.

Por ser uma pesquisa realizada anualmente, um outro viés possível nos dados que precisa ser investigado é a presença anual dos países. Quantos países presentes no ano anterior estão ausentes no seguinte e vice-versa? Isso pode ser observado no gráfico abaixo:

Gráfico, Gráfico de linhas

Descrição gerada automaticamente Figura 9 – Países adicionados e removidos ao longo do tempo

É notável que, dado que o primeiro ano dos dados é 2005, o ano seguinte foi o com mais países adicionados, mas também com muitos ausentes. Entre 2012 e 2019 o número de países se manteve bastante estável, mas em 2020, possivelmente devido a pandemia, a pesquisa não foi realizada em um número razoável de países, que voltaram a aparecer em 2021.

**6.2 Distribuição do *score* de felicidade**

A distribuição do total de registros do score deve ser feita com a pergunta feita aos entrevistados para a sua obtenção: “sendo 1 a pior vida possível e 10 a melhor, onde você está agora?”. A forma da pergunta faz com que resultados entre 1 e 2, assim como 9 e 10, sejam improváveis, visto que corresponderiam as piores e melhores vidas possíveis respectivamente. O entrevistado é induzido à uma resposta entre 2 e 9, o que pode ser visto no gráfico abaixo.

Gráfico, Histograma

Descrição gerada automaticamente Figura 10 – Distribuição do score no dataset completo

Além disso, outra informação relevante é que a média geral dos dados (5.47) não possui uma variação significativa ao longo do tempo.

Tabela

Descrição gerada automaticamente com confiança média Figura 11 – Score ao longo do tempo

É necessário destacar o ano de 2020 da figura acima que, ainda que esteja apresentando uma melhora no score de felicidade mesmo em meio a uma pandemia, é preciso lembrar que 2020 teve uma grande queda no número de países presentes na pesquisa, o que certamente distorceu os dados para esse ano. Tal situação prejudica consideravelmente a hipótese de extrair informações do impacto da pandemia no score.

No gráfico abaixo podemos ver a média geral do score para cada país e visualizar que existem diferenças consideráveis entre as regiões do mundo para o score.

Mapa

Descrição gerada automaticamente Figura 12 – Média do score por país. Para melhorar a visualização não foi utilizada a escala 1-10 devido à indução de respostas menos extremas pelo formato da pergunta realizado.

As diferenças visíveis entre o *score* das diferentes regiões do mundo mostram que talvez, mesmo com a diferença e dados entre as regiões, seja possível classificar a região do mundo e estimar de forma bastante assertiva o score caso a região esteja sendo considerada no modelo de regressão. Podemos destacar ainda mais as diferenças entre as regiões com a figura abaixo.

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente Figura 13 – Gráfico de caixas do score separado por região do mundo.

**6.3 Relações entre as métricas**

A primeira etapa na exploração realizada nos dados para entender as relações entre os atributos foi o mapa de calor representando as correlações positivas e negativas entre elas.

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente Figura 14 – Mapa de calor das correlações entre os atributos.

É possível identificar na figura acima que os atributos *gdp*, *social\_support* e *hle* possuem as maiores correlações positivas com o score, enquanto *corruption* e *negative\_affect* possuem as maiores negativas. Ainda que não se possa tirar conclusões de causalidade de um mapa de correlações, elas fazem sentido intuitivamente, mostrando que os dados da pesquisa estão dentro do esperado. Todavia, as correlações apresentadas são do conjunto inteiro dos dados de todas as pesquisas. Devido a isso, faz-se necessário uma compreensão da evolução dos valores ao longo do tempo, o que pode ser observado nas figuras 15 e 16 abaixo.

Interface gráfica do usuário

Descrição gerada automaticamente com confiança baixa Figura 15 – Atributos ao longo do tempo com ajuste do intervalo do eixo y.

A diferença a figura 15 acima e a 16 abaixo está na escala escolhida do eixo y. A figura 15 apresenta as escalas normalizadas, permitindo a comparação entre os scores enquanto a figura 16 apresenta os dados ao longo do tempo com o intervalo do eixo y ajustado individualmente para melhor visualização da evolução e do intervalo de confiança.

**Gráfico

Descrição gerada automaticamente com confiança média** Figura 16 – Atributos ao longo do tempo com eixo y ajustado individualmente.

É destacável, mais uma vez, a diferença dos valores na maior parte dos atributos para o ano de 2020, indicando que a observação dos impactos da pandemia no score deve ser feita apenas entre os países presentes na pesquisa em 2020, não podendo ser feita com o conjunto inteiro dos dados.

Por fim, vemos abaixo a distribuição dos valores de cada atributo, demarcando a sua respetiva média e mediana e escalados individualmente.

Gráfico, Histograma

Descrição gerada automaticamente Figura 17 – Distribuição dos valores dos atributos numéricos da pesquisa.

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

Buscando validar a viabilidade do objetivo principal e objetivos secundários apresentados anteriormente, fez-se necessário realizar etapas de preparação dos dados e testes preliminares de diferentes algoritmos de regressão selecionados.

**7.1 *Feature engineering***

Utilizando o mesmo padrão *Commands* utilizado na etapa de processamento inicial dos dados, foram realizadas as seguintes transformações:

1. Categorização das regiões em valores numéricos
2. Categorização dos países em valores numéricos. É importante salientar que este atributo não foi utilizado para os modelos preliminares e não deve ser utilizado nos modelos em produção referentes ao objetivo principal desse trabalho, pois ocasionaria *overfitting*.
3. Criar atributo contendo o arredondamento do valor do score para o inteiro mais próximo. O objetivo é facilitar regressões mais simples para estimar intervalos dos dados nas extremidades onde não existem dados informados de score – valores entre 1 e 2 e entre 9 e 10.
4. Criar atributo contendo os valores de HLE escalados entre 0 e 1. Como esse valor corresponde à idade humana ele varia em uma escala consideravelmente maior que os outros, fazendo-se necessária uma normalização. Para tal foi utilizado o atributo criado no item 3 para uma regressão de quais seriam os limites máximos e mínimos para o HLE, utilizando esses valores para efetuar uma normalização min-max.

**7.2 Modelos lineares de regressão**

Com o propósito de comparar diferentes modelos lineares de regressão foram escolhidos os seguintes algoritmos:

* *Regressão Linear.* O mais tradicional e matematicamente simples dos algoritmos de regressão. Tem o objetivo de minimizar a soma residual dos quadrados das diferenças entre os dados observados no dataset e os dados preditos pela aproximação linear.
* *Cross Validation Elastic Net.* Regressão linear que realiza a seleção de variáveis, permitindo uma maior precisão para dados esparsos e/ou altamente correlacionados. Além disso, realiza a busca dos coeficientes através de validação cruzada.
* *Regressão Bayesiana.* Regressão linear utilizando modelo probabilístico.

Para realizar o teste foi feita a separação simples entre dados para treino e teste, onde são selecionados aleatoriamente 20% dos dados para teste e 80% dos dados para treino. Tal separação não será utilizada para o modelo final, conforme será visto adiante, mas apenas para os testes preliminares. Os modelos foram validados utilizando o Coeficiente de Determinação R², medida estatística da proximidade entre os dados e a linha de regressão.

|  |  |
| --- | --- |
| **Modelo** | **R²** |
| Regressão Linear | 0.76179 |
| Elastic Net CV | 0.74117 |
| Regressão Bayesiana | 0.76172 |

Tabela 5 – Score dos algoritmos de regressão linear

A regressão linear teve o melhor resultado dentre os algoritmos utilizados, conforme imagem abaixo.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente Figura 18 – Resultado do ajuste da Regressão Linear

Para a criação dos modelos os atributos País e Ano foram descartados e o parâmetro Região foi utilizado em sua forma categórica numérica.

**7.3 Modelos não-lineares de regressão**

Utilizando os mesmos critérios de validação (R²), atributos e separação dos dados dos modelos lineares, foram selecionados os seguintes algoritmos não lineares de classificação:

* *SVR.* Algoritmo que utiliza a classificação SVM, mas ao invés de classificar o item, utiliza a curva resultante para determinar resultados para os dados desconhecidos.
* *KNN.* Compara dados semelhantes entre si com o objetivo de inferir uma determinada característica sobre um dado desconhecido.
* *Decision Tree Regressor*. Variação do algoritmo de classificação por árvore de decisão, mas que utiliza a variância como medida de impureza ao invés da entropia. Os dados são separados no espaço através dos nodos e a regressão pode ser, dependendo da parametrização, a média dos valores presentes naquele específico ramo.
* *Random Forest.* Treina várias árvores de decisão, obtidas a partir de amostras do dataset e realiza predições utilizando a média dos valores obtidos.

|  |  |
| --- | --- |
| **Modelo** | **R²** |
| SVR | 0.89038 |
| KNN | 0.88596 |
| Árvore de Decisão | 0.89785 |
| Floresta aleatória | 0.81456 |

Tabela 6 – Score dos algoritmos de regressão não-linear

Dentre os algoritmos de regressão não-linear, a que obteve melhor score foi o de Floresta Aleatória, mas todos eles tiveram resultados substancialmente melhores que as regressões lineares.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente Figura 19 – Resultado do ajuste da Floresta Aleatória

Conforme mostra a imagem acima, o *Random Forest* pareceu ter um bom resultado. No entanto uma das grandes desvantagens conhecidas desse algoritmo é o risco de *overfitting*. Devido a isso é necessário utilizar uma separação de dados e teste diferente para validar os modelos.

**7.4 Implementação da validação cruzada**

Uma das formas mais utilizadas para separação de treino e teste na criação de modelos é a validação cruzada. Nela são selecionados N amostras aleatórias dos dados e o treinamento com o algoritmo selecionado é feito N vezes independentes, cada uma delas com uma das amostras ausentes e sendo usada como teste. A precisão do modelo é medida pela média do score resultante de cada treinamento, permitindo a visualização do score de cada amostra e a identificação de vieses dos dados.

Todavia a forma com que as funções prontas de separação de treino e teste presentes nas bibliotecas utilizadas no trabalho funcionam não suprem a necessidade do formato do trabalho, visto que a pesquisa é separada anualmente e qualquer novo dado futuro para ela será também um conjunto de dados para um próximo ano. Devido a tal necessidade, foi construído uma validação cruzada customizada, onde os dados foram divididos anualmente e cada rodada de treinamento utiliza um ano diferente para teste. O resultado permitiu uma comparação mais precisa entre os algoritmos não-lineares utilizados.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ano** | **SVR** | **KNeighborsRegressor** | **RandomForestRegressor** | **DecisionTreeRegressor** |
| 2005 | 0.683112 | 0.863846 | 0.725676 | 0.489011 |
| 2006 | 0.718683 | 0.877169 | 0.867135 | 0.734891 |
| 2007 | 0.866629 | 0.836069 | 0.853888 | 0.775121 |
| 2008 | 0.885824 | 0.879394 | 0.893776 | 0.771842 |
| 2009 | 0.817139 | 0.872427 | 0.837847 | 0.745531 |
| 2010 | 0.894297 | 0.913976 | 0.907515 | 0.829120 |
| 2011 | 0.856254 | 0.841392 | 0.869580 | 0.748563 |
| 2012 | 0.917136 | 0.912196 | 0.919316 | 0.841234 |
| 2013 | 0.919534 | 0.911238 | 0.890768 | 0.822141 |
| 2014 | 0.913898 | 0.917598 | 0.893309 | 0.791457 |
| 2015 | 0.901501 | 0.917798 | 0.919257 | 0.825505 |
| 2016 | 0.947316 | 0.930745 | 0.919498 | 0.843015 |
| 2017 | 0.945479 | 0.941714 | 0.921948 | 0.810442 |
| 2018 | 0.962358 | 0.945720 | 0.911967 | 0.813279 |
| 2019 | 0.947619 | 0.888779 | 0.906640 | 0.876215 |
| 2020 | 0.863148 | 0.874042 | 0.912703 | 0.813790 |
| 2021 | 0.982706 | 0.932067 | 0.964478 | 0.935178 |

Tabela 7 – Score dos algoritmos de regressão não-linear utilizando validação cruzada anual

A tabela 7 acima mostra que o KNN foi o algoritmo mais estável dentre os selecionados, enquanto na tabela 8 vemos que ele foi com a maior média e menor desvio padrão, mesmo não possuindo o maior *score* da amostra.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SVR** | **KNeighborsRegressor** | **RandomForestRegressor** | **DecisionTreeRegressor** |
| μ | 0.883684 | 0.897422 | 0.889135 | 0.792137 |
| σ | 0.081029 | 0.033655 | 0.051668 | 0.092566 |
| Mín. | 0.683112 | 0.836069 | 0.725676 | 0.489011 |
| Máx. | 0.982706 | 0.945720 | 0.964478 | 0.935178 |

Tabela 8 – Informações sobre o score dos algoritmos utilizados na validação cruzada anual.

Dessa forma, através dessa validação cruzada, foi possível identificar que o KNN é um modelo suficientemente adequado para o objetivo principal deste projeto.

**7.5 Validando a importância da região**

Com o propósito de validar a hipótese de que é possível classificar um registro nas regiões do mundo, foi realizado um teste de regressão utilizando uma validação cruzada semelhante à vista acima, mas ao invés de utilizar o ano, foi utilizada a região para separar as amostras. Se o *score* da regressão fosse baixo isso identificaria que a região do mundo possui um papel relevante na classificação, mas se continuasse um *score* satisfatório, significaria pouco impacto da região do mundo na subjetividade da resposta e possivelmente a hipótese seria falsa.

|  |  |  |
| --- | --- | --- |
| **Região** | **R²** | **R² Ajustado** |
| asia | 0.150213 | 0.128424 |
| eastern europe | 0.088633 | 0.049644 |
| northern africa | -0.601122 | -0.873653 |
| sub-saharan africa | -0.393935 | -0.418498 |
| latin amer. and carib | -0.613331 | -0.656353 |
| c.w. of ind. states | -0.542212 | -0.613528 |
| oceania | -2.000113 | -3.143013 |
| western europe | 0.244880 | 0.223534 |
| near east | 0.587021 | 0.566627 |
| northern america | -7.821380 | -10.889687 |
| baltics | -0.097171 | -0.334397 |

Tabela 9 – Resultado da validação cruzada por região.

A tabela acima mostra que a regressão não performou bem separando os dados por região, o que indica o impacto da região do mundo na definição do score.

# 8. Links

Os arquivos, códigos, artefatos e análises mais detalhadas realizadas no projeto podem ser encontradas 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/>>