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
Junho 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

O campo de Inteligência Artificial e Aprendizado de Máquina tem sido cada vez mais utilizado por organizações dos setores público e privado. 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.

Beneficiando-se dessa interdisciplinaridade, o presente trabalho utilizará o Relatório de Felicidade Mundial, uma pesquisa realizada a partir de uma iniciativa da Organização das Nações Unidas (ONU), conhecida como Rede de Soluções para o Desenvolvimento Sustentável. O propósito do trabalho 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 intuito de facilitar a análise da felicidade social como indicador de desenvolvimento, a ONU, 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.*

Em 2011, a ONU recomendou aos países participantes que utilizassem indicadores de felicidade para acompanharem o seu desenvolvimento e em 2012 o relatório foi lançado, sendo atualizado anualmente e tendo seus 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, têm-se como objetivos secundários analisar a existência de alterações nos valores causadas 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 e *K-Nearest Neighbors* (KNN) serão utilizados para o objetivo principal. 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 e Redes Neurais 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* 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, dada a sua grande popularidade no uso para aplicações semelhantes e o amplo número de bibliotecas de código aberto que implementam os algoritmos propostos. Algumas das bibliotecas utilizadas são: *sklearn*, *keras*, *numpy* e *pandas*.

Todo o código-fonte e o processo de análise serão disponibilizados na plataforma GitHub, e os resultados poderão ser acessados via REST API.

3. Canvas Analítico

Com o intuito 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[1], adaptando o modelo para as necessidades do projeto. São utilizadas as seguintes etapas do Canvas: questões, fontes de dados, hipóteses (como substituto de heurísticas), validação e implementação. Nas seções a seguir é apresentada a definição de cada uma dessas etapas com seu respectivo papel para realização do projeto.

---

[1] *Software Analytics Canvas*, <https://www.feststelltaste.de/software-analytics-canvas>.

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 estipuladas no presente trabalho baseiam-se no objetivo principal e objetivos 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 obtido?
3. A pandemia causou algum impacto nos índices obtidos de felicidade mundial?

3.2 Fontes de dados

Conforme introduzido na seção 2 e detalhado na seção 4, são utilizados os dados disponibilizados no Relatório de Felicidade Mundial de 2021, realizado pela iniciativa de Rede de Soluções para o Desenvolvimento Sustentável 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, baseada 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, essa percepção pode ser influenciada pela própria qualidade de vida descrita nas métricas, pois, dado um longo período, as pessoas podem se acostumar com a situação 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, a 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 etapas. 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 etapas acima descritas não são concluídas necessariamente de forma sequencial, visto que os resultados de uma etapa podem afetar o desenvolvimento da etapa 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, tendo como propósito a criação de um índice de desenvolvimento que não fosse exclusivamente atrelado a métricas financeiras, como o PIB, e 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: o primeiro contendo os dados até 2020 e o segundo 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 destes para análise e modelagem.

Nas Tabelas 1 e 2, são listados os atributos dos dois *datasets* utilizados. Como a maioria dos atributos está presente nos dois arquivos, a Tabela 1 contém todos os atributos utilizados de 2021 e a Tabela 2 corresponde aos dados até 2020, apenas com o que está ausente nos dados de 2021.

Tabela 1 – Atributos do *dataset* com os 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	Decimal

	resultado é a média dessas respostas.	
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

Fonte: Autor (2022).

Tabela 2 – Atributos do *dataset* com os dados até 2020.

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

Fonte: Autor (2022).

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. A Tabela 3 apresenta o *dataset* com as colunas utilizadas no tratamento.

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

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

Fonte: Autor (2022).

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 pode ser observado na Figura 1.

Figura 1 – world-happiness-report/src/data_preparation_commands.py

```
 9 Commands() \
10     .add_command(CleanColumns) \
11     .add_command(RegionJoin) \
12     .add_command(Affects) \
13     .add_command(DatasetsJoin) \
14     .add_command(MissingData) \
15     .add_command(RegionCleaning) \
16     .execute()
17
```

Fonte: Autor (2022).

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.

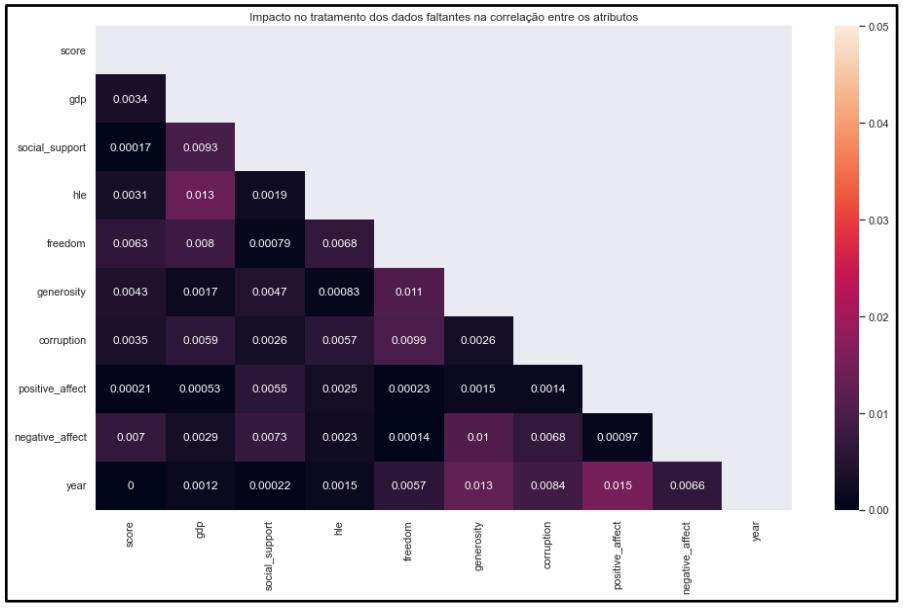
A terceira etapa visa lidar com duas colunas ausentes nos dados de 2021 presentes nos dados históricos: “positive affects” e “negative affects”. Ambos os atributos são responsáveis por quantificar o estado sentimental das pessoas nos respectivos países. Devido às questões recentes no cenário mundial 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.

A quinta etapa tem como propósito preencher os dados faltantes nos *datasets* com o menor impacto possível nas relações entre os atributos. 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.

Figura 2 - Mapa de calor baseado no valor absoluto de alteração das correlações entre as variáveis após o tratamento dos dados faltantes.



Fonte: Autor (2022).

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.

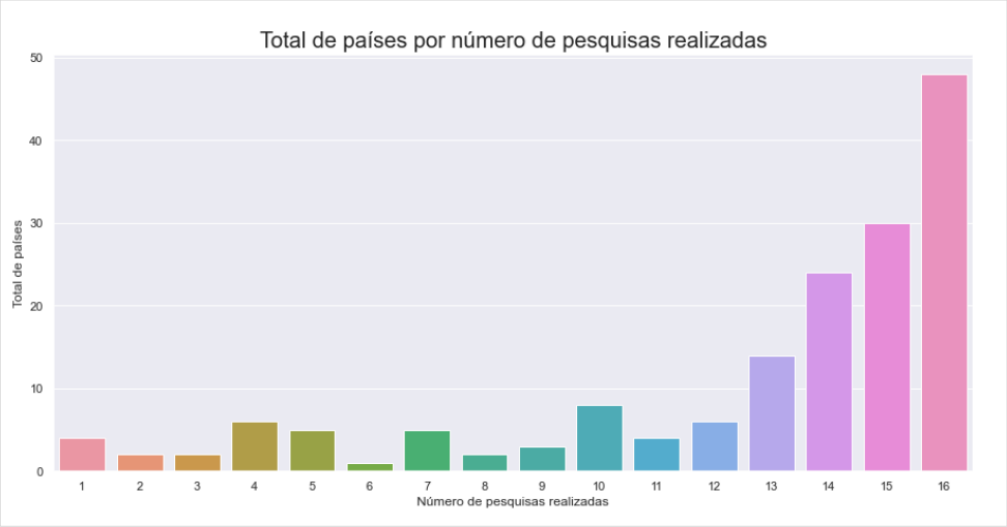
6. Análise e Exploração dos Dados

A exploração dos dados teve como principal objetivo identificar possíveis vieses e desbalanceamentos e validar hipóteses. 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 os demais 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 uma edição. No entanto, a presença ao longo do tempo não foi constante, tendo apenas aproximadamente 29% dos países com registros em todas as edições.

Figura 3 - Presença total dos países na pesquisa.

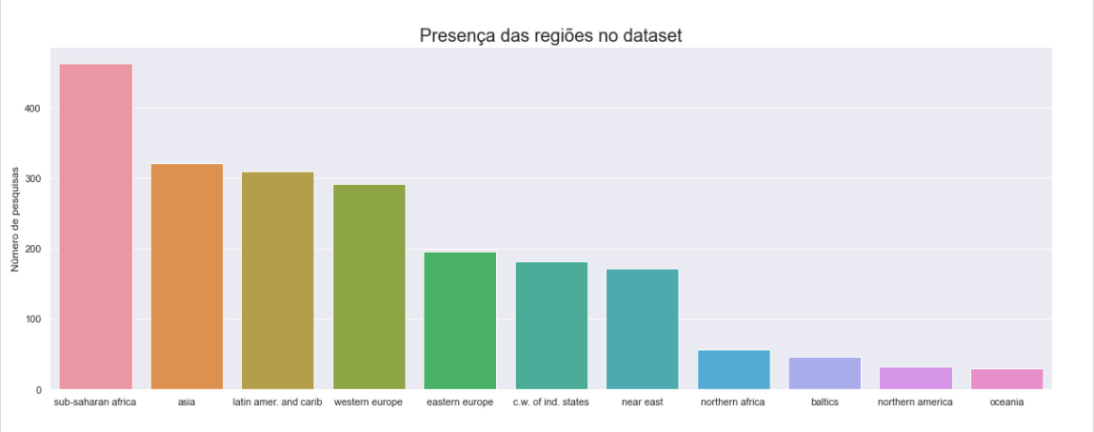


Fonte: Autor (2022).

O gráfico presente na Figura 3 destaca a presença de países nas pesquisas realizadas. É notável que aproximadamente 70% dos países participaram em mais de 12 edições da pesquisa. Contudo, o número de países com baixa presença deve ser levado em consideração na etapa de validação dos modelos de aprendizagem.

Uma análise semelhante pode ser realizada para as regiões do mundo destacando uma informação importante: 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.

Figura 4 – Presença das regiões do mundo na pesquisa. Note que regiões como a América do Norte possuem baixa presença pois são compostas por poucos países.

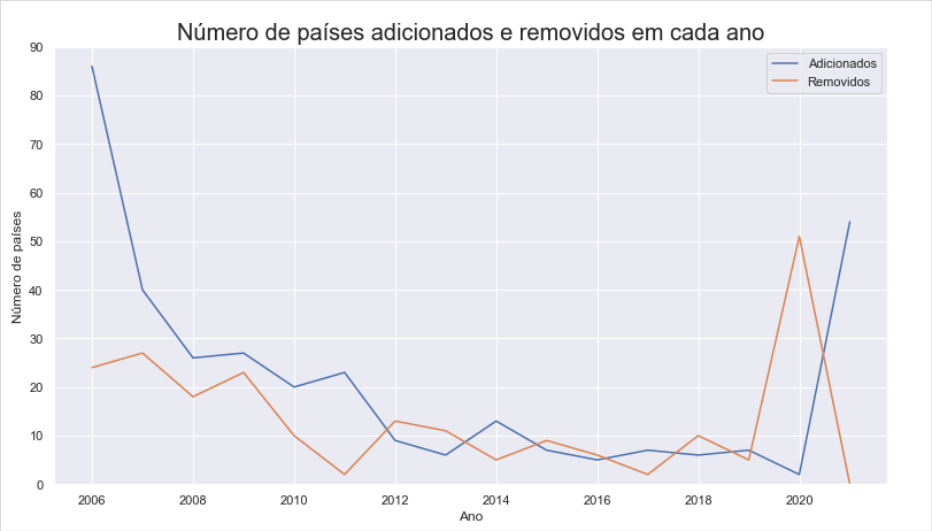


Fonte: Autor (2022).

A discrepância apontada pela Figura 4 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.

Dada a realização anual da pesquisa, é relevante considerar o conjunto de países presentes a cada ano, ou seja, os países presentes em um determinado ano, mas ausentes no ano seguinte e vice-versa.

Figura 5 – Países adicionados e removidos ao longo do tempo.



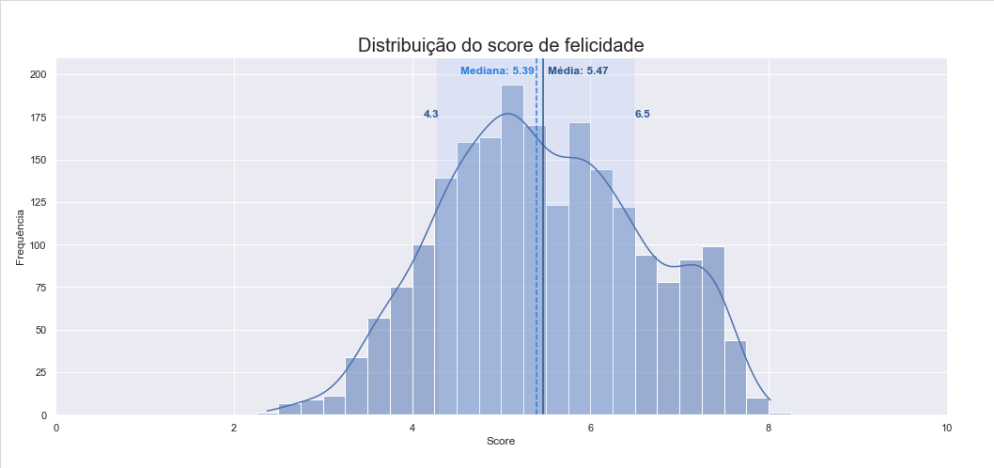
Fonte: Autor (2022).

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 a partir da pergunta realizada aos entrevistados: “sendo 1 a pior vida possível e 10 a melhor, onde você está agora?”. A estrutura da pergunta faz com que resultados 1 e 2, assim como 9 e 10, sejam improváveis, visto que correspondem às piores e melhores vidas possíveis, respectivamente. O entrevistado é então influenciado à uma resposta entre 2 e 9, o que pode ser observado na Figura 6.

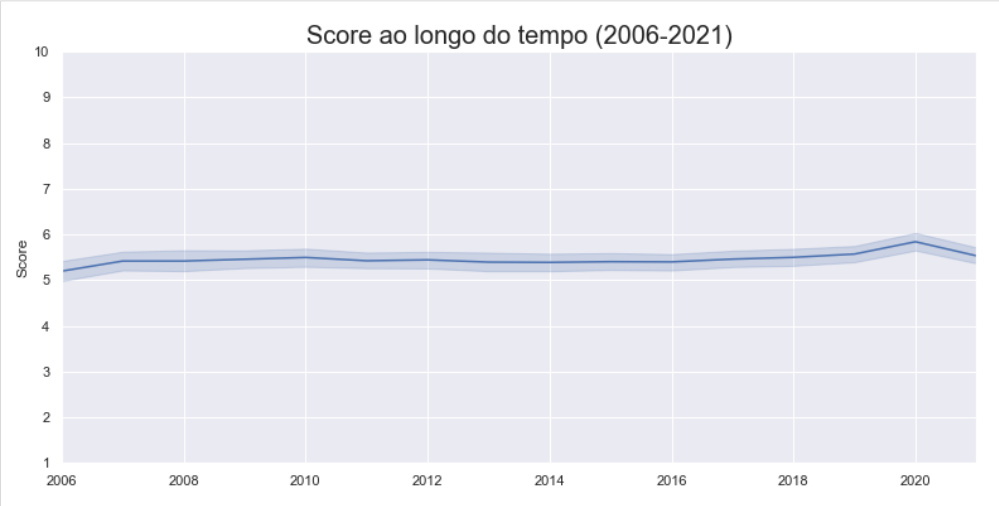
Figura 6 – Distribuição do *score* no *dataset* completo.



Fonte: Autor (2022).

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

Figura 7 – Score ao longo do tempo.

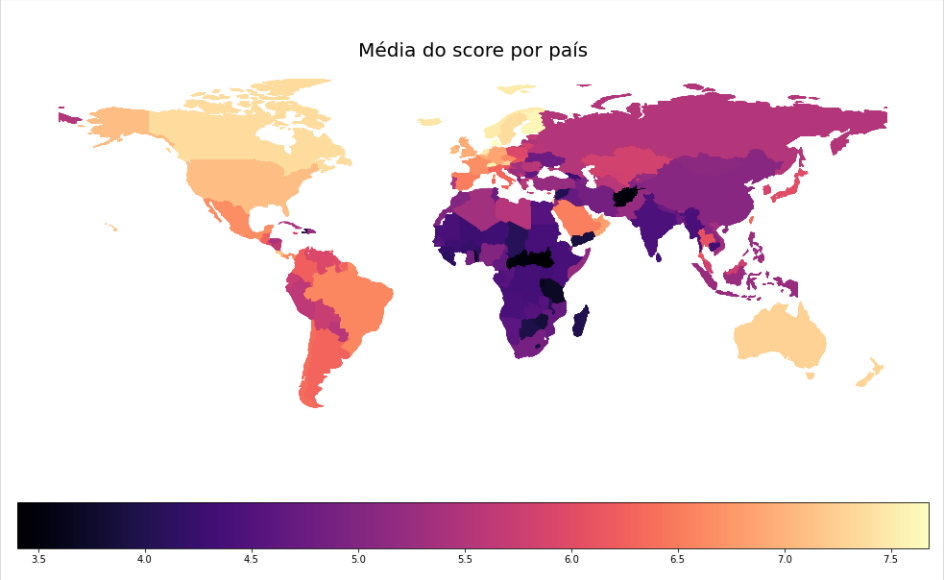


Fonte: Autor (2022).

Na Figura 7, é necessário destacar o ano de 2020, pois ainda que apresente uma melhora no *score* de felicidade mesmo em meio a uma pandemia global, é preciso salientar que em 2020 houve uma grande queda no número de países presentes na pesquisa, como mostra a Figura 5. Isto leva a uma inconsistência com os dados obtidos nesse ano, prejudicando a análise da hipótese de extrair informações do impacto da pandemia no *score*.

Na Figura 8, podemos ver a média geral do score para cada país e visualizar que existem diferenças consideráveis entre as regiões do mundo para o score.

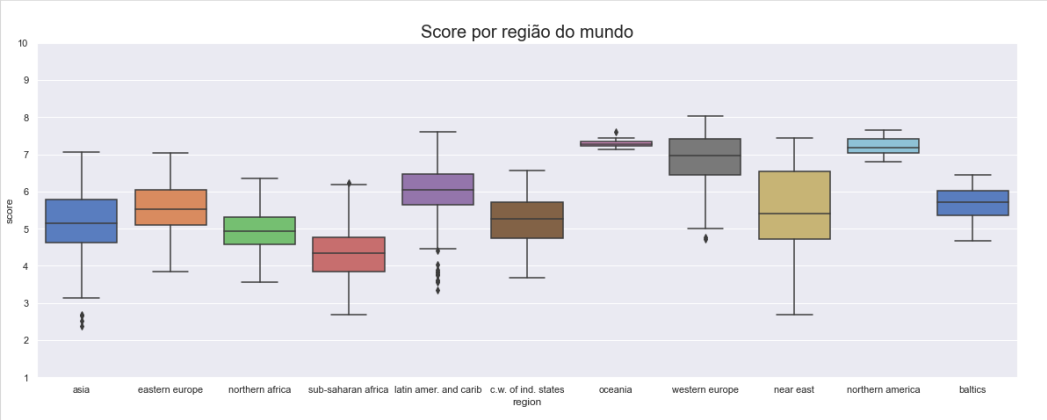
Figura 8 – Média do score por país. Para melhorar a visualização não foi utilizada a escala 1-10 devido à indução de respostas menos extremas pelo formato da pergunta realizada.



Fonte: Autor (2022).

As diferenças visíveis entre o *score* das diferentes regiões do mundo indica a possibilidade de classificar o dado por região do mundo, estimando de forma assertiva o *score* caso a região esteja sendo considerada no modelo de regressão.

Figura 9 – Gráfico de caixas do score separado por região do mundo.

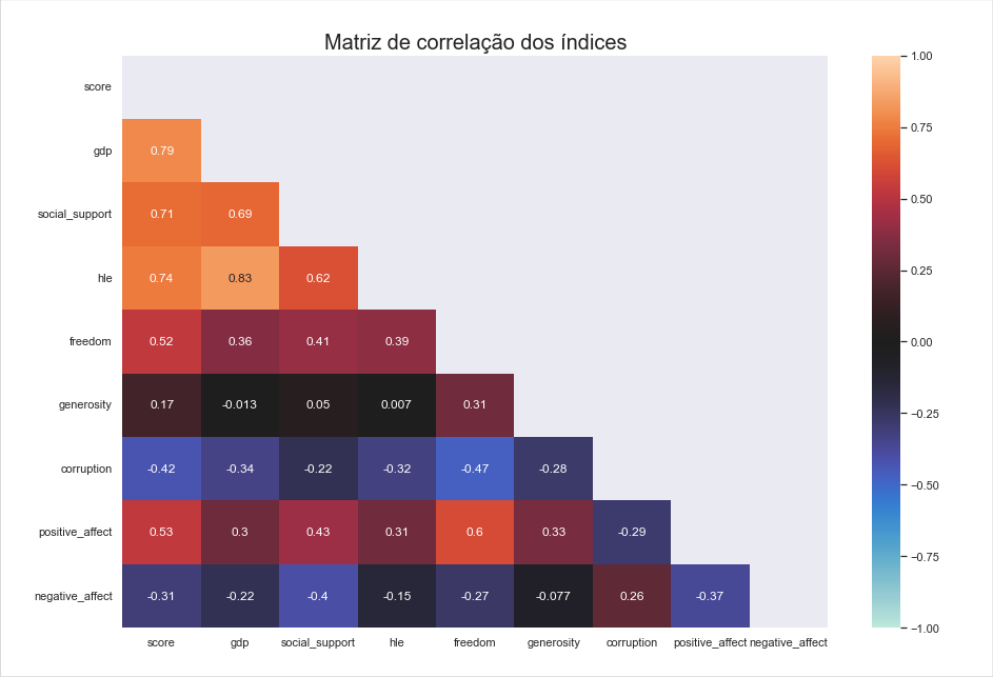


Fonte: Autor (2022).

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.

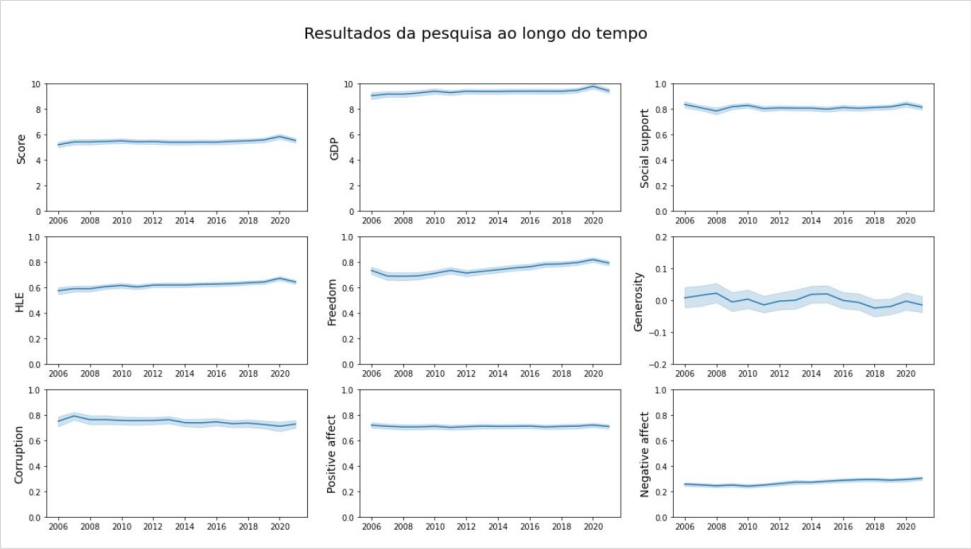
Figura 10 – Mapa de calor das correlações entre os atributos.



Fonte: Autor (2022).

É possível identificar 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. Assim, faz-se necessário uma compreensão da evolução dos valores ao longo do tempo.

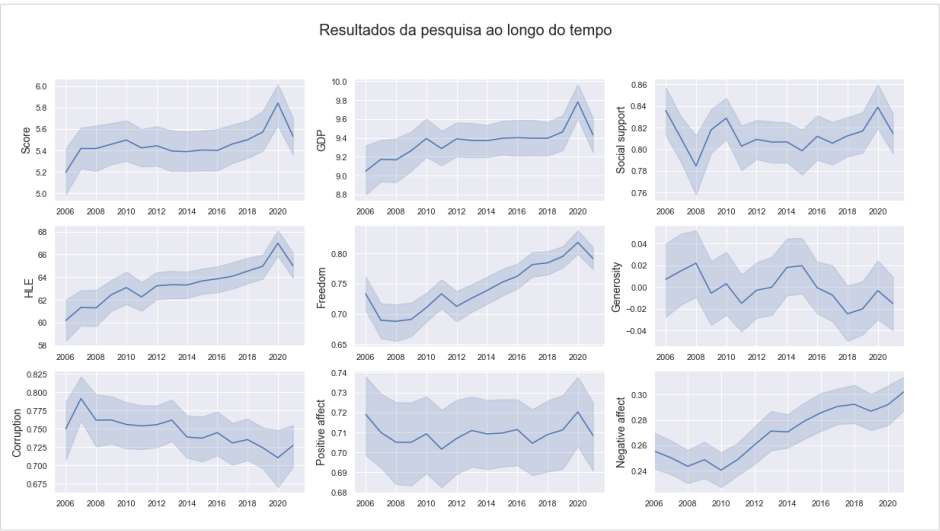
Figura 11 – Atributos ao longo do tempo com ajuste do intervalo do eixo y.



Fonte: Autor (2022).

A Figura 11 e a 12 estão na escala escolhida do eixo y. A Figura 11 apresenta as escalas normalizadas, permitindo a comparação entre os scores enquanto a Figura 12 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.

Figura 12 – Atributos ao longo do tempo com eixo y ajustado individualmente.



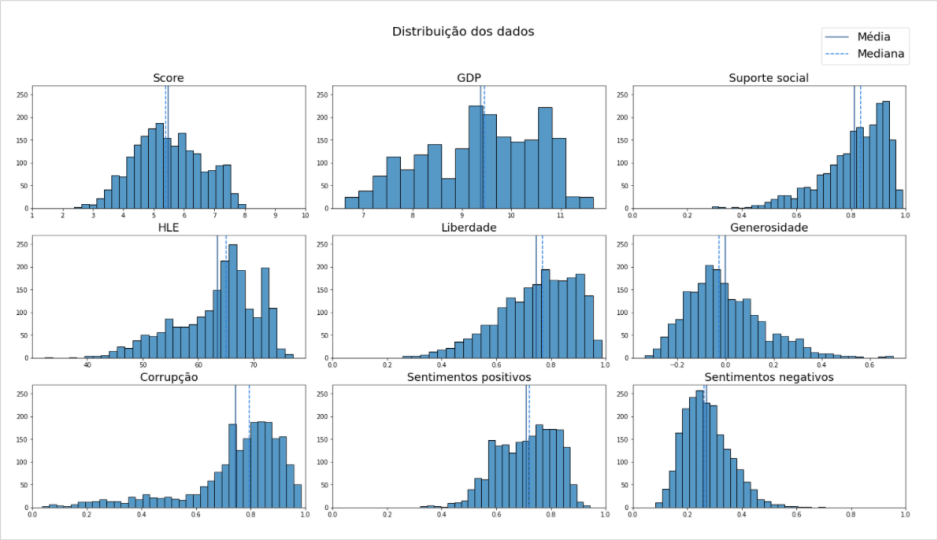
Fonte: Autor (2022).

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 respectiva média e mediana e escalados individualmente.

Figura 13 – Distribuição dos valores dos atributos numéricos da pesquisa.



Fonte: Autor (2022).

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

Buscando confirmar a viabilidade dos objetivos do presente trabalho, 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;
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 diferente, fazendo-se necessária uma normalização. Para isto, 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 mín-máx.

7.2 Modelos lineares de regressão

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

- Regressão Linear[2];
- *Elastic Net* com *Cross Validation*[3];
- Regressão Bayesiana[4].

Para realizar a comparação, foi feita uma separação simples dos dados em treino e teste, onde são selecionados aleatoriamente 80% dos dados para treino e 20% para treino. Tal separação não será utilizada para o modelo final, apresentado na seção 7.4, apenas para os testes preliminares. Os modelos foram validados utilizando o Coeficiente de Determinação R^2, medida estatística da proximidade entre os dados e a linha de regressão.

Tabela 4 – Score dos algoritmos de regressão linear.

Modelo	**R^2**
Regressão Linear	0.76179

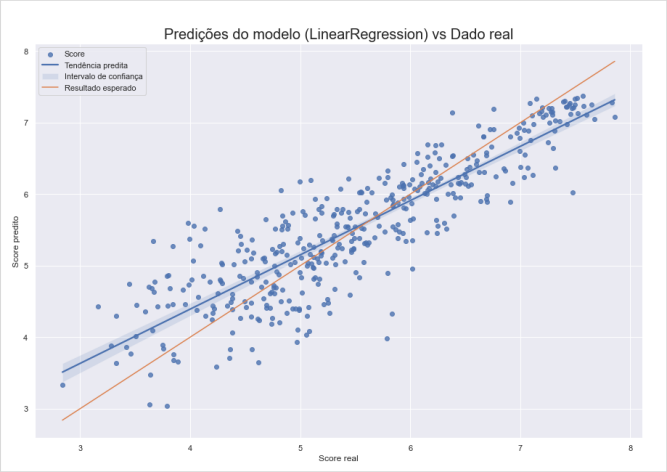
[2] Documentação LinearRegression, <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression>.
[3] Documentação ElasticNetCV, <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNetCV.html>.
[4] Documentação BayesianRidge, <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html#sklearn.linear_model.BayesianRidge>.

Elastic Net CV	0.74117
Regressão Bayesiana	0.76172

Fonte: Autor (2022).

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

Figura 14 – Resultado do ajuste da Regressão Linear.



Fonte: Autor (2022).

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 (R^2), atributos e separação dos dados dos modelos lineares, foram selecionados os seguintes algoritmos não lineares de classificação:

- *Support Vector Regression* (SVR)[5];
- *K-Nearest Neighbors Regressor* (KNN)[6];

[5] Documentação SVR, <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>.
[6] Documentação *KNeighborsRegressor*, <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>.

- *Decision Tree Regressor*[7];
- *Random Forest Regressor*[8].

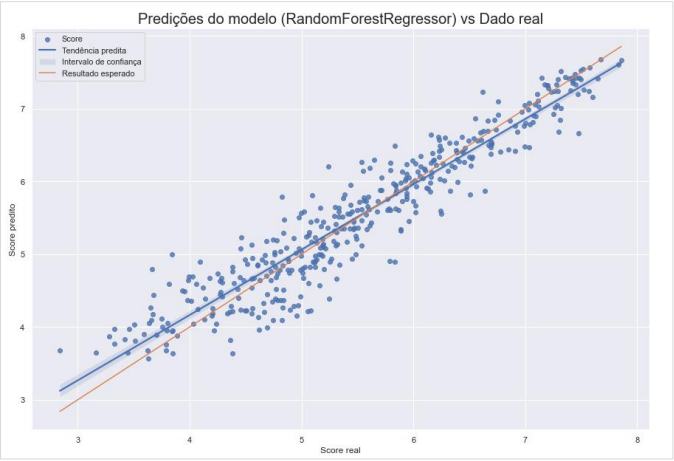
Tabela 5 – Score dos algoritmos de regressão não-linear.

Modelo	**R^2**
SVR	0.89038
KNN	0.88596
Árvore de Decisão	0.89785
Floresta aleatória	0.81456

Fonte: Autor (2022).

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

Figura 15 – Resultado do ajuste da Floresta Aleatória.



Fonte: Autor (2022).

Conforme mostra a Figura 15, o *Random Forest* demonstrou ter um bom resultado. No entanto, uma das grandes desvantagens conhecidas desse algoritmo é

[7] Documentação *DecisionTreeRegressor*, <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>.
[8] Documentação *RandomForestRegressor*, <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>.

o risco de *overfitting*. Assim, é essencial realizar a separação dos dados para teste de forma a minimizar este problema.

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[9]. 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.

Tendo em vista a realização anual da pesquisa, a inclusão de informações será sempre um conjunto de dados referentes a um novo ano. Diante disso, foi construída uma validação cruzada onde cada amostra corresponde aos dados de um ano presente na pesquisa. Em cada rodada de treinamento, uma das amostras é utilizada para testar o respectivo modelo. O resultado, então, permitiu uma comparação mais precisa entre os algoritmos não-lineares utilizados.

Na Tabela 6 vemos que ele demonstrou a maior média e menor desvio padrão, mesmo não possuindo o maior *score* máximo.

Tabela 6 – Informações sobre o score dos algoritmos utilizados na validação cruzada anual.

	SVR	**KNN**	**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

Fonte: Autor (2022).

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

[9] *Cross-validation*, <https://scikit-learn.org/stable/modules/cross_validation.html>.

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.

Tabela 7 – Resultado da validação cruzada por região.

Região	**R^2**	**R^2 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

Fonte: Autor (2022).

A Tabela 7 mostra que a regressão não performou bem separando os dados por região, indicando impacto da região 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 em 29 de março de 2022, <https://worldhappiness.report/>