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
Agosto 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
Agosto de 2022

SUMÁRIO

1. Introdução

A Inteligência Artificial tem sido cada vez mais utilizada por organizações dos setores público e privado. 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 sucesso da área.

Beneficiando-se desse avanço, o presente trabalho utiliza 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, este trabalho tem por objetivo principal construir um modelo de Aprendizado de Máquina que seja capaz de, dado as métricas utilizadas no relatório, estimar o valor do índice de felicidade em um determinado país. Além disso, têm-se como objetivos secundários criar modelos de classificação para 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 e analisar a existência de alterações nos valores causadas pela pandemia (2020 e 2021).

Foram utilizados algoritmos de aprendizado de máquina para regressão, classificação e clusterização dos dados. Os algoritmos de regressão foram utilizados para a previsão do *score* de felicidade, os algoritmos de classificação para a previsão

da região do mundo de um registro e os de clusterização foram utilizados na etapa de análise exploratória dos dados. Os dois melhores modelos de regressão e classificação foram publicados em produção e disponibilizados via API.

A linguagem escolhida para desenvolver os modelos de Aprendizado de Máquina juntamente com as análises e transformações de dados necessárias foi 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.

Todo o código-fonte do processo de análise, implementação dos modelos e a API estão na plataforma GitHub. Os resultados podem ser acessados via API ou visualizados em uma página web desenvolvida com Angular e cujo código-fonte também está disponível no GitHub.

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.

3.1 Questões

A etapa de questões corresponde às perguntas feitas sobre os dados para obter informações relevantes na identificação e resolução dos problemas apontados por eles. As questões estipuladas no presente trabalho baseiam-se nos objetivos principal e secundários apresentados. São elas:

---

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

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 ONU.

3.3 Hipóteses

A terceira etapa, originalmente chamada de “Heurísticas” foi adaptada ao problema do projeto. Foi definida como o conjunto das 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 de um país. Porém, essa percepção varia de acordo com a região em que o entrevistado está. Dessa forma, seria possível descobrir a região do mundo de um país apenas baseado na relação entre as métricas da pesquisa com o índice de felicidade.

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.

4. Coleta de Dados

Conforme mencionado anteriormente, os dados foram obtidos por meio do Relatório de Felicidade Mundial. Contendo dados anuais entre 2005 e 2021 coletados 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 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

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 dos dados realizado se deu em cinco etapas, cujo objetivo final era obter um *dataset* único e consistente, sem dados faltantes e com os dados

corretamente preenchidos. Cada etapa se propõe a resolver um problema encontrado nos *datasets*, sendo executadas sequencialmente. A implementação dessas etapas utiliza uma adaptação do padrão de design *Commands*[2]. A chamada às etapas de processamento pode ser vista na Figura 1.

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

```
16      def ingest(self) -> None:
17          self.download_data()
18          _ = execute_data_combination()
19          _ = execute_data_preparation()
20          _ = execute_feature_engineering()
21          _ = execute_data_augmentation()
22
```

Fonte: Autor (2022).

A primeira etapa realiza a ingestão dos dados, fazendo download dos arquivos a partir de links configurados via variáveis de ambiente.

A segunda etapa combina o *dataset* dos dados históricos com os dados de 2021. Para tal, realiza a remoção de colunas desnecessárias, preenchimento da região do mundo dos países nos dados históricos com a informação de 2021 e ajuste dos dados faltantes de "*positive affects*" e "*negative affects*" com a atribuição dos dados de 2020 em 2021 e a realização da média entre o ano anterior e o posterior do país aos faltantes no *dataset* histórico.

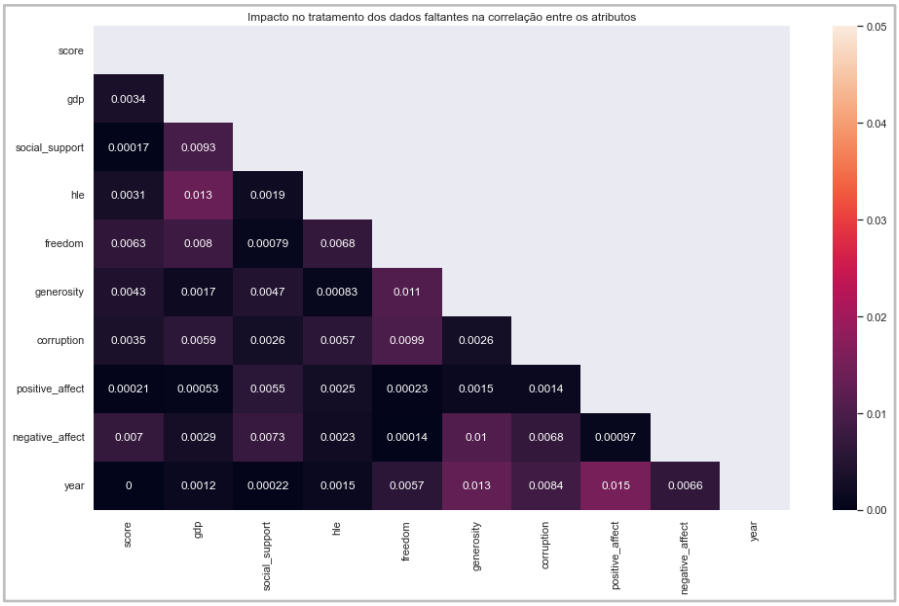
A terceira etapa tem como propósito preencher os dados faltantes nos *datasets* com o menor impacto possível nas relações entre os atributos. Devido à baixa quantidade de dados faltantes, o algoritmo escolhido para preencher os dados foi:

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);

[2] Command, <https://refactoring.guru/pt-br/design-patterns/command>.

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).

Na quarta etapa é realizado o processo chamado de *feature engineering*, que cria atributos baseados nos dados existentes. São criadas colunas para categorizar numericamente países e regiões, normalização dos dados de HLE e a criação de um *dataset* separado apenas com os países presentes na pesquisa de 2020, para uma melhor análise do impacto da pandemia.

A quinta e última etapa utiliza o método SMOTE[3] para balancear os dados por região do mundo. Como algumas regiões possuem mais países que outras, sua representatividade no *dataset* é menor. Buscando resolver isso essa etapa cria um *dataset* balanceado por região a ser utilizado no algoritmo de classificação. No

[3] SMOTE, <https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html>.

entanto, para uma melhor validação dos resultados do modelo, é realizada a separação de 15% dos dados antes do balanceamento com o propósito de obter métricas de avaliação do modelo de classificação mais próximas a novos dados reais.

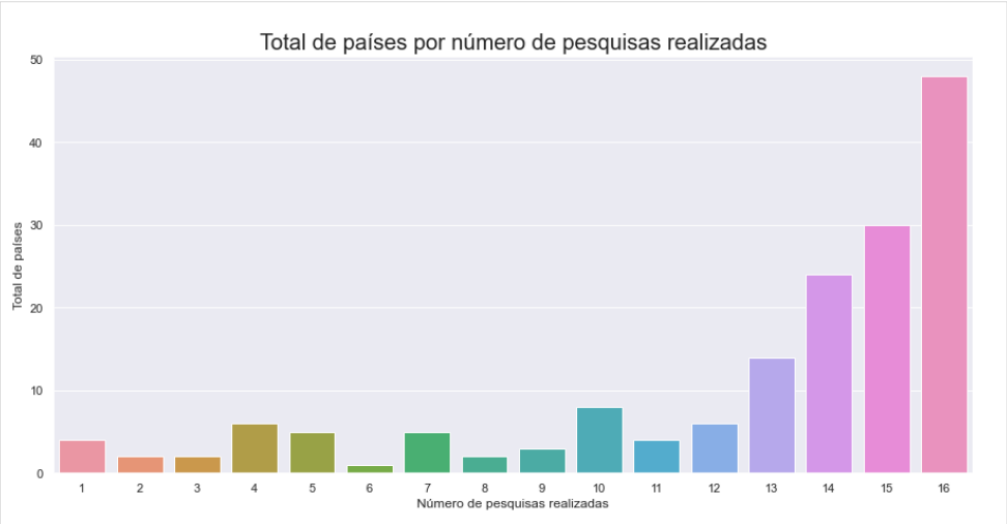
6. Análise e Exploração dos Dados

A exploração dos dados teve como principal objetivo identificar vieses, 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.



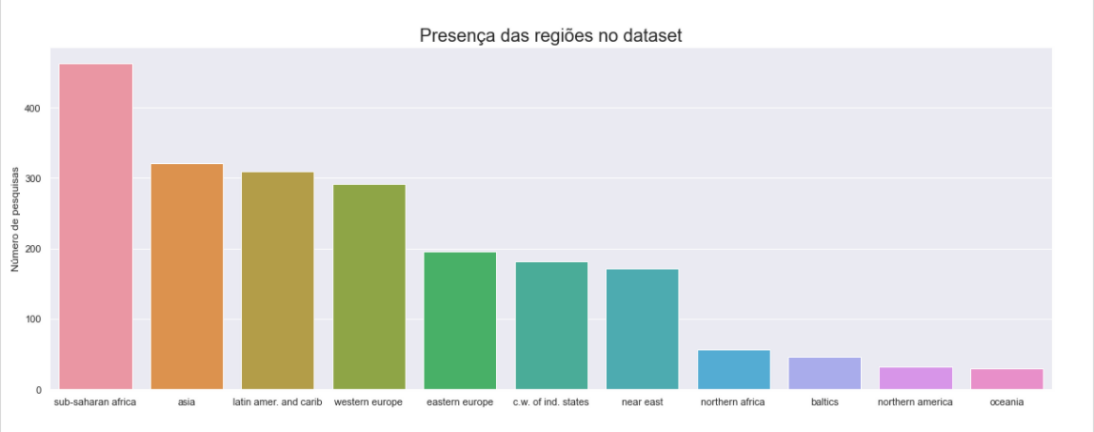
Fonte: Autor (2022).

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.

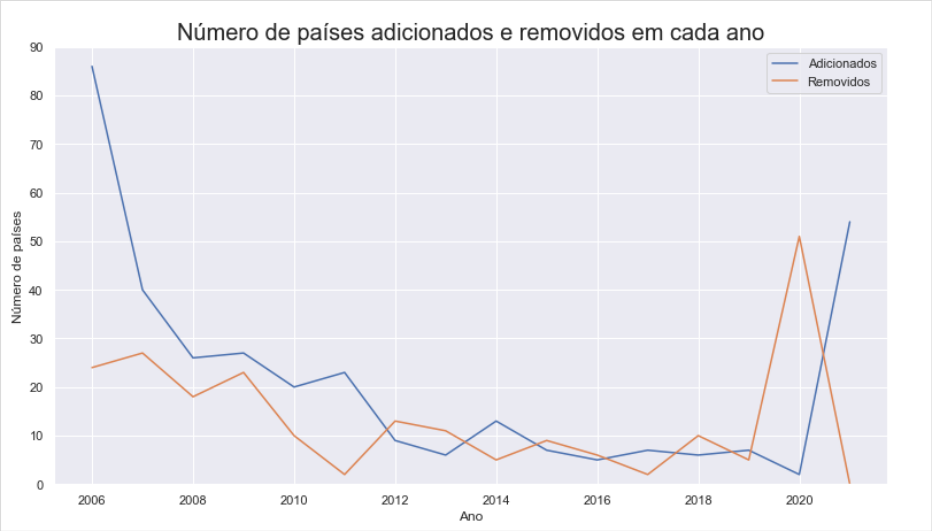


Fonte: Autor (2022).

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 e por isso a necessidade do balanceamento mencionados na seção 5.

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.



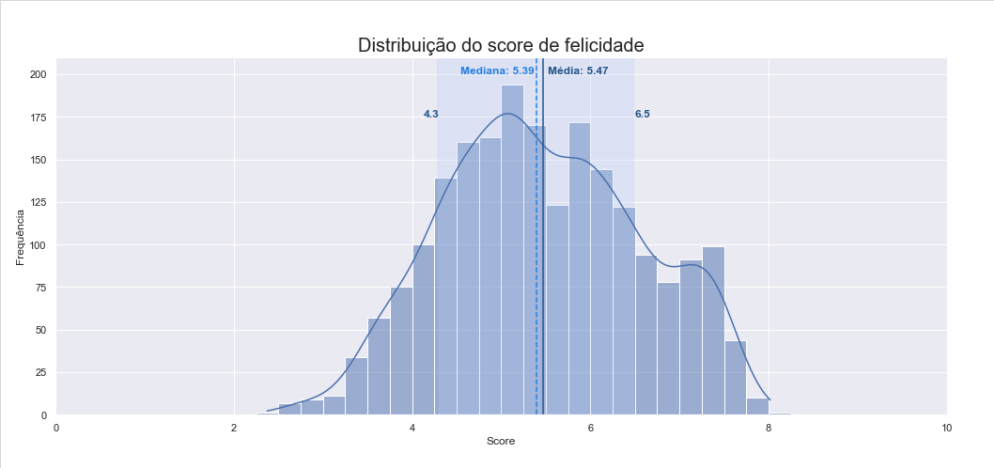
Fonte: Autor (2022).

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. Dessa forma, qualquer análise do impacto da pandemia deve ser feita considerando apenas os países presentes em 2020.

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.

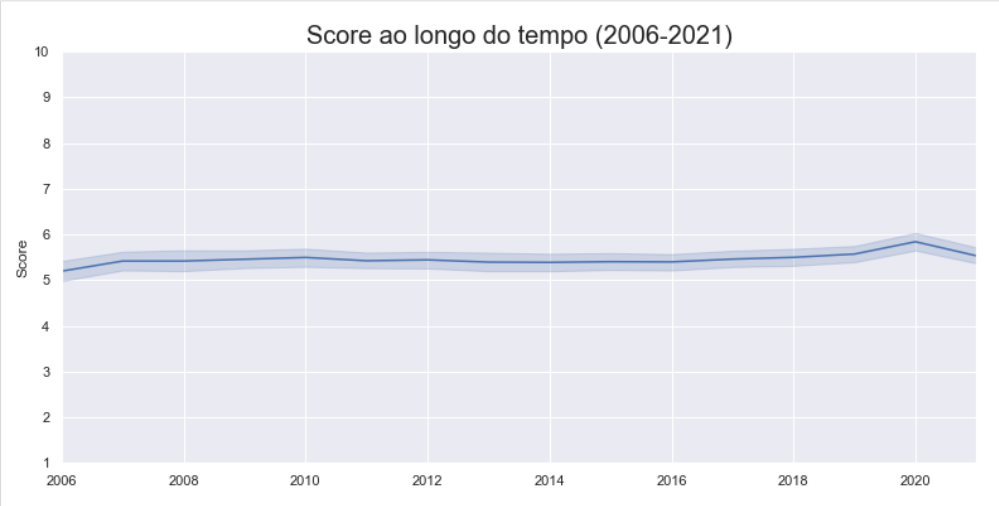
Figura 6 – Distribuição do *score* no *dataset* completo.



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.

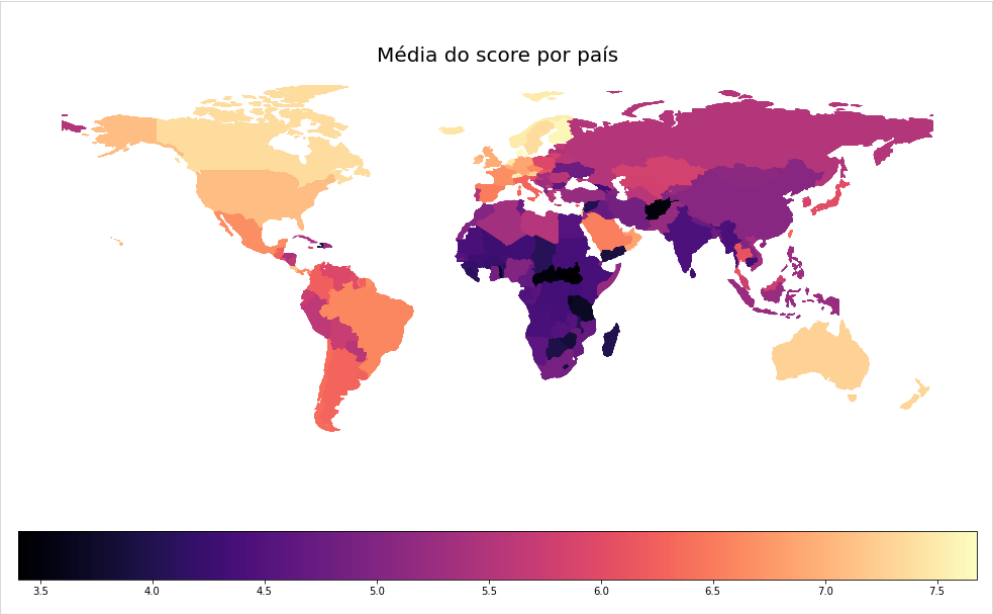
Figura 7 – Score ao longo do tempo.



Fonte: Autor (2022).

Na Figura 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.

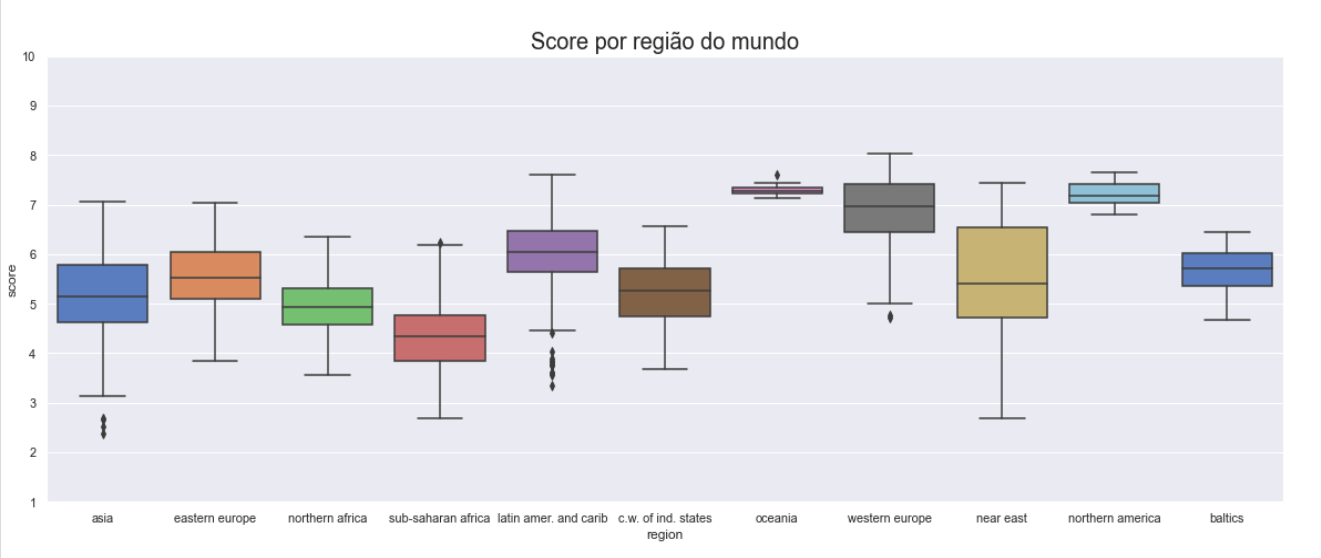
Figura 8 – Média do score por país. Para melhorar a visualização não foi utilizada a escala 1-10 devido à indução de respostas menos extremas pelo formato da pergunta realizada.



Fonte: Autor (2022).

O comportamento do *score* no mapa da Figura 8 indica a viabilidade da hipótese de classificação por região do mundo e a necessidade de considerar a região no modelo de regressão, o que pode ser visto na Figura 9.

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

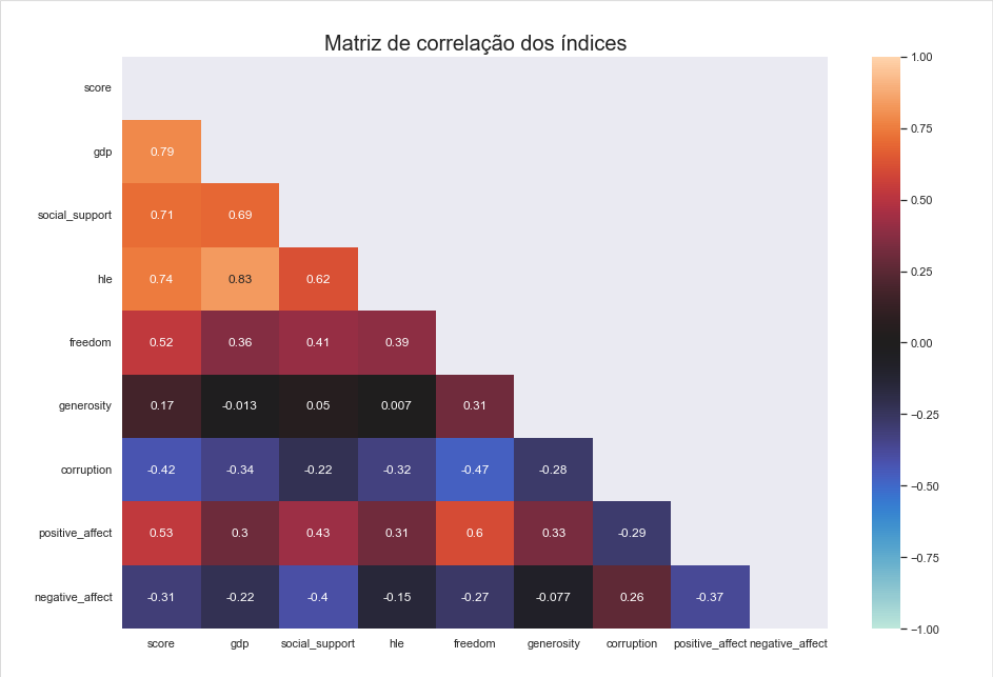


Fonte: Autor (2022).

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.

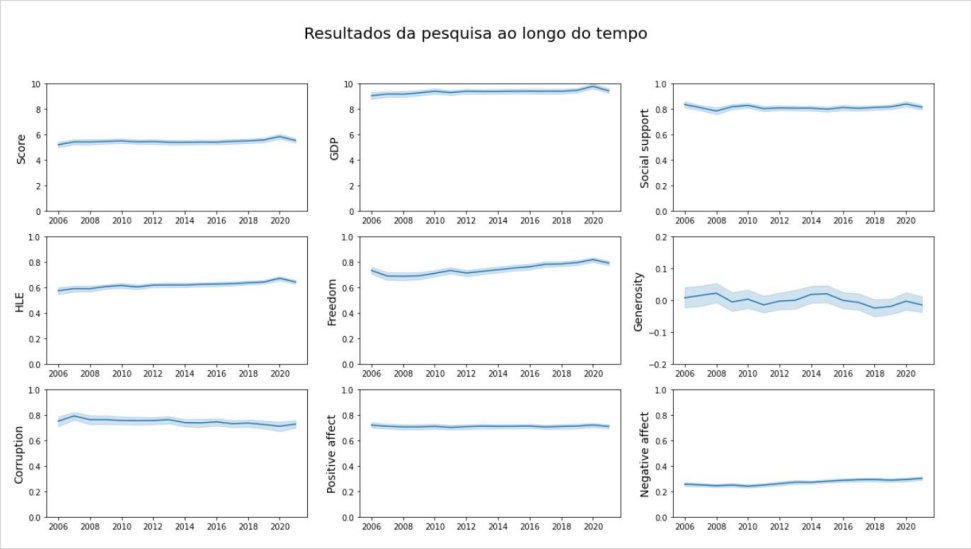
Figura 10 – Mapa de calor das correlações entre os atributos.



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.

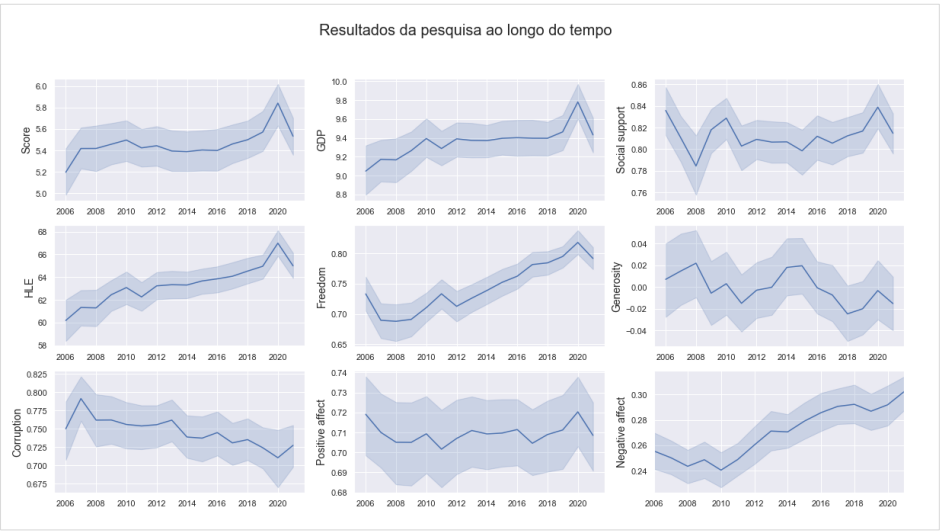
Figura 11 – Atributos ao longo do tempo com ajuste do intervalo do eixo y.



Fonte: Autor (2022).

A Figura 11 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.



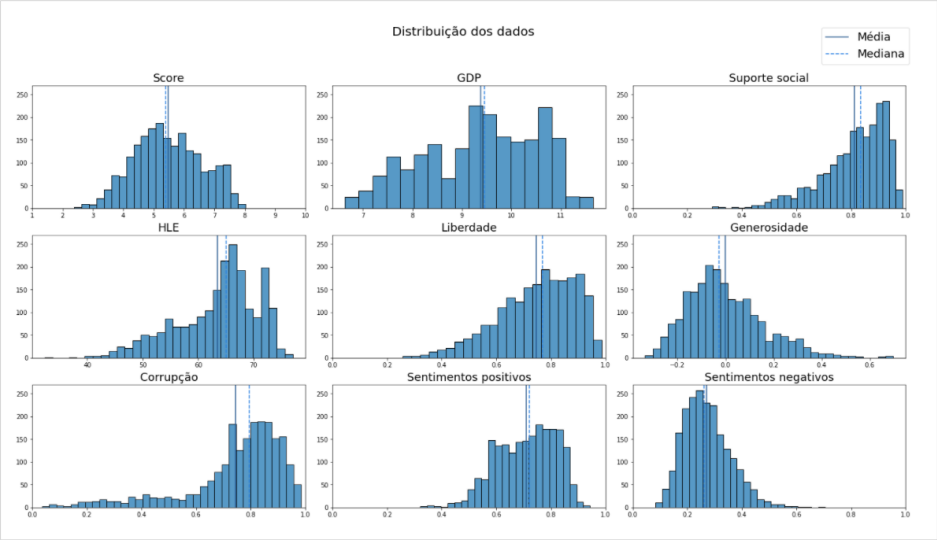
Fonte: Autor (2022).

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).

6.4 Impacto da pandemia

Devido ao grande número de países ausentes na pesquisa de 2020, a análise do impacto da pandemia ficou bastante limitada. Dado o histórico apenas dos países presentes no respectivo ano não houve alterações significativas por causa da pandemia e a regressão do *score* para o ano de 2020 não apresentou dificuldades. Apenas os valores de *Negative Affect* e *Generosity* apresentaram variação relevante em 2020, mas sem alterar as relações entre os atributos. Dessa forma, é possível afirmar que os dados do *dataset* escolhido não são suficientes para uma análise do impacto social da pandemia na felicidade das pessoas.

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*

Conforme mencionado na seção 5, foi realizada a criação de novos atributos a partir dos dados existentes. Destes, destaca-se a criação de um atributo contendo os valores de HLE escalados entre 0 e 1. Como esse valor corresponde à idade humana, ele varia em uma escala diferente. Para resolver isso, foi utilizada a normalização min-máx. No entanto, como não existem registros de *score* com valores nos intervalos [1, 2] e [9, 10], fez-se uma regressão simples de HLE para esses valores de *score* para utilizar na normalização.

7.2 Modelos lineares de regressão

Com o propósito de comparar diferentes modelos lineares de regressão foram escolhidos os algoritmos de Regressão Linear[4], *Elastic Net* com *Cross Validation*[5] e Regressão Bayesiana[6].

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
Elastic Net CV	0.74117

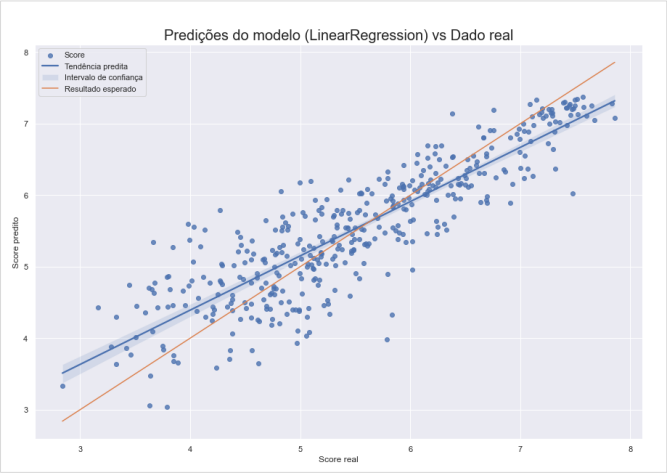
[4] Documentação LinearRegression, <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression>.
[5] Documentação ElasticNetCV, <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNetCV.html>.
[6] Documentação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

Regressão Bayesiana	0.76172

Fonte: Autor (2022).

A regressão linear teve o melhor resultado dentre os algoritmos utilizados, conforme a Figura 14.

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 algoritmos não lineares, sendo eles o *Support Vector Regression* (SVR)[7], *K-Nearest Neighbors Regressor* (KNN)[8], *Decision Tree Regressor*[9] e *Random Forest Regressor*[10].

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

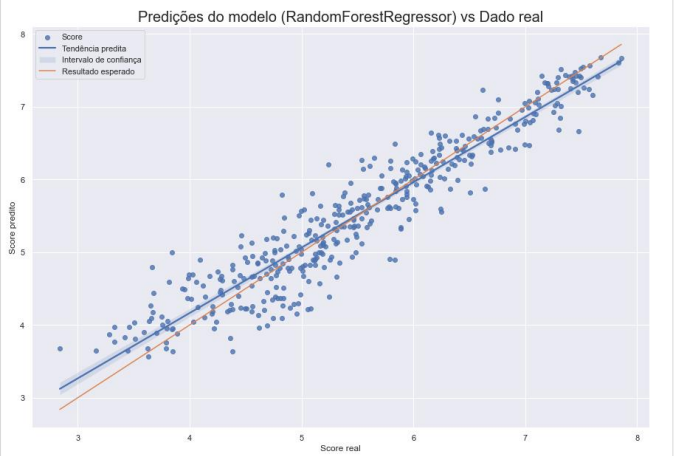
Tabela 5 – Score dos algoritmos de regressão não-linear.

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, o que obteve melhor score foi o *Random Forest*, mas todos eles apresentaram resultados melhores que as regressões lineares.

Figura 15 – Resultado do ajuste da Floresta Aleatória.



Fonte: Autor (2022).

Segundo apresentado pela Figura 15, o *Random Forest* demonstrou ter um bom resultado. No entanto, uma das grandes desvantagens conhecidas desse algoritmo é o risco de *overfitting*[11]. Assim, é essencial realizar a separação dos dados para teste de forma a minimizar este problema.

7.4 Implementação da validação cruzada

[11] Underfitting e Overfitting, <https://didatica.tech/underfitting-e-overfitting/>.

Uma das formas mais utilizadas para separação de treino e teste na criação de modelos é a validação cruzada[12]. 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.

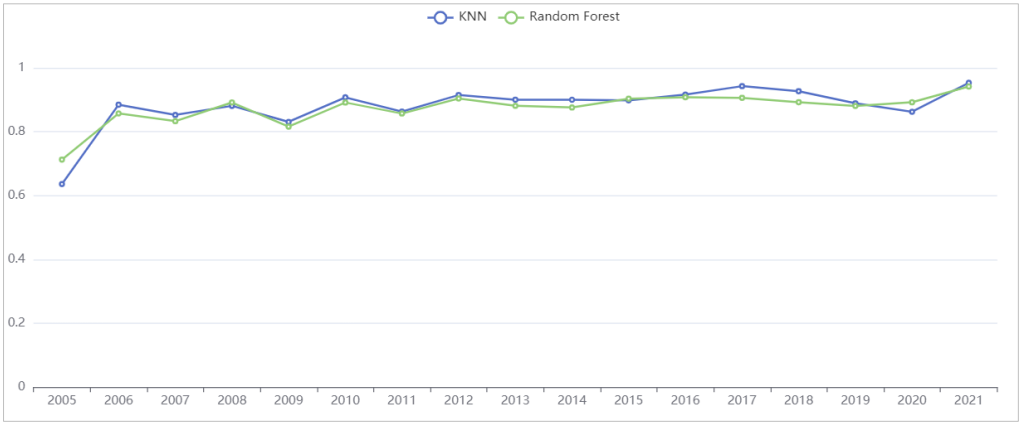
Tabela 6 – Informações sobre o score dos algoritmos utilizados na validação cruzada anual.

	SVR	**KNN**	**RandomForestRegressor**	**DecisionTreeRegressor**
μ	0.883684	0.897422	0.889135	0.792137
σ	0.081029	0.033655	0.051668	0.092566

Fonte: Autor (2022).

Diante desses resultados, os algoritmos *Random Forest* e KNN podem ser considerados suficientemente adequados para a implementação em produção.

Figura 16 – Comparação do R^2 entre KNN e *Random Forest*.



Fonte: Autor (2022).

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

7.5 Validando a importância da região

Buscando 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 R^2 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.

7.6 Modelo de classificação da região

Devido ao bom desempenho dos algoritmos *Random Forest* e KNN para a regressão do *score*, além da sua boa explicabilidade, optou-se por utilizá-los também para validar o modelo de classificação por região do mundo. Foram utilizados os dados balanceados, mencionados na seção 5, para o treinamento do modelo e os dados

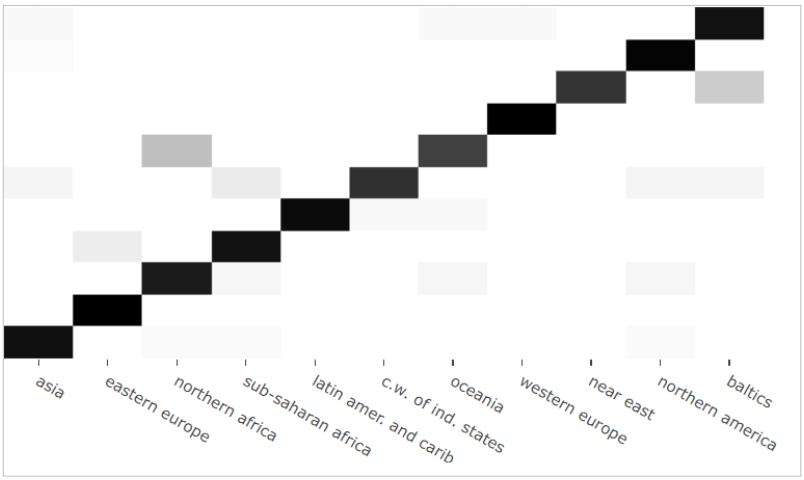
separados para validação para aferir a qualidade dos resultados. A performance do Random Forest se mostrou superior ao KNN, apresentando os seguintes resultados:

Tabela 8 – Resultado da classificação por região.

Região	**Precisão**	**Recall**
asia	0.9375	0.9375
eastern europe	0.8710	0.9310
northern africa	0.6667	0.7500
sub-saharan africa	0.9583	0.9857
latin amer. and carib	1.0000	0.9565
c.w. of ind. states	0.8889	0.8889
oceania	1.0000	0.8000
western europe	0.9535	0.9318
near east	0.9545	0.8077
northern america	0.8333	1.0000
baltics	0.7778	1.0000

Fonte: Autor (2022).

Figura 17 – Matrix de confusão da classificação por região com *Random Forest.*



Fonte: Autor (2022).

8. Aplicação de modelos de Aprendizado de Máquina

Com o objetivo de realizar a publicação dos modelos acessíveis via API publicamente, tendo os seus resultados exibidos em uma página web consumindo essa API, com o menor custo de infraestrutura possível, decidiu-se por realizar a

publicação utilizando os serviços gratuitos da Azure[13]. O serviço de API realiza, em sua inicialização, a ingestão e tratamento dos dados. Após isso é possível chamar os *endpoints* referentes aos modelos de classificação e regressão, obtendo dados sobre suas respectivas performances e permitindo a realização de previsões com os modelos. O consumo desses *endpoints* pode ser realizado através da documentação no Swagger, cujo link está disponibilizado na seção 11.

Ambos os códigos do *frontend* e da API estão automatizados através de GitHub Actions[14] que publicam uma imagem do Docker[15] no serviço de Container Registry da Azure[16] para todas as alterações realizadas na *branch* principal. Quando essa imagem é atualizada, o container é publicado no Serviço de Aplicativo da Azure[17]. Todos os dados apresentados neste trabalho sobre os modelos e tratamentos dos dados podem ser obtidos na API, sendo alguns deles visualizáveis na aplicação web, como a importância dos atributos para cada modelo utilizado. O treinamento e execução dos modelos são também realizáveis sob demanda via API. Além disso, toda a documentação necessária para leitura e execução do código-fonte está disponível nos arquivos README.md presentes nas pastas raiz dos repositórios utilizados.

8.1 Parâmetros dos modelos utilizados

Após as diversas análises e comparações realizadas na etapa de avaliação dos modelos, que podem ser observadas com mais detalhes nos arquivos da pasta *analysis* no código-fonte, os quatro modelos selecionados (dois de regressão e dois de classificação) foram parametrizados da seguinte forma:

---

[13] Devido ao uso dos serviços de menor custo, a memória RAM alocada para essas publicações é de apenas 1GB. Dessa forma é possível que as publicações apresentem alguma lentidão na execução.
[14] GitHub Actions, <https://github.com/features/actions>.
[15] Docker, <https://www.docker.com/>.
[16] Azure Container Registry <https://azure.microsoft.com/en-us/services/container-registry/>.
[17] Azure App Service <https://azure.microsoft.com/pt-br/services/app-service/>.

- *RandomForestRegressor:* parametrização padrão da biblioteca exceto pelos parâmetros *random_state*, com o valor 42, e o parâmetro *max_depth,* com o valor 10;
- *KNeighborsRegressor:* foi utilizada a parametrização padrão da biblioteca;
- *KNeighborsClassifier:* foi utilizada a parametrização padrão da biblioteca;
- *RandomForestClassifier:* parametrização padrão da biblioteca exceto pelo parâmetro *random_state*, com o valor 42;

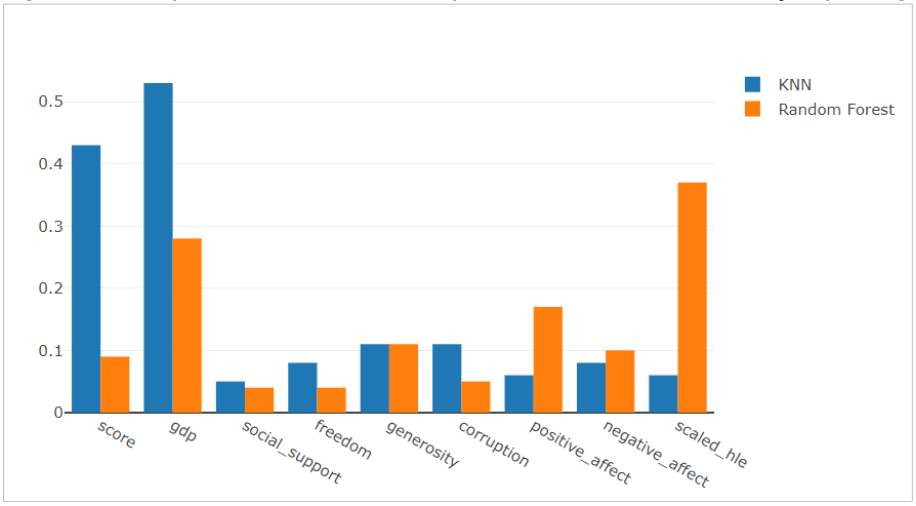
Os outros algoritmos utilizados durante o desenvolvimento do projeto foram testados com diversas combinações de parâmetros e não foram implementados na versão final da API pelo seu desempenho inferior aos algoritmos escolhidos. Ainda assim, as implementações e avaliações desses modelos podem ser verificadas no repositório do código. Os resultados dos modelos aplicados em produção podem ser vistos na seção 7.

8.2 Importância dos atributos e explicabilidade do modelo

Buscando seguir a boa prática de construção de modelos de aprendizado de máquina explicáveis[18] com o propósito de evitar ou, ao menos, estar ciente dos possíveis vieses dos modelos utilizados, a API implementada em produção disponibiliza a importância de cada atributo para os algoritmos utilizados. Os resultados podem ser vistos nas Figuras 18, 19 e 20.

[18] Explainable AI: extraindo explicações e aumentando a confiança dos modelos de ML, <https://medium.com/data-hackers/explainable-ai-extraindo-explica%C3%A7%C3%B5es-e-aumentando-a-confian%C3%A7a-dos-modelos-de-ml-3a89b7b5a584>.

Figura 18 – Importância das variáveis para o modelo de classificação por região.



Fonte: Autor (2022).

Figura 19 – Importância das variáveis para o modelo de regressão do *score* para cada cluster da validação cruzada - KNN.

Fonte: Autor (2022).

Figura 20 – Importância das variáveis para o modelo de regressão do *score* para cada cluster da validação cruzada – *Random Forest*.

Fonte: Autor (2022).

9. Discussão dos resultados

Diante dos resultados obtidos pelo presente projeto, observou-se a possibilidade de, utilizando algoritmos de baixa complexidade e com resultados facilmente explicáveis via *feature importances*[19], prever o índice de felicidade de um país baseado nas métricas dos dados da pesquisa. Identificou-se também que existe diferença na subjetividade da percepção de felicidade dentre as várias regiões do mundo, de forma que permitiu a classificação com boa precisão da região do mundo para um determinado conjunto de métricas e seu respectivo índice de felicidade.

10. Conclusão

Dado os resultados apresentados neste trabalho, a interação entre pesquisas sociológicas e a área de Inteligência Artificial mostrou-se promissora. Sendo a percepção de felicidade algo bastante individual e cultural, a possibilidade de identificar padrões em valores subjetivos e gerar inteligência a partir dos mesmos abre caminho para aplicações computacionais cada vez mais preocupadas com o desenvolvimento humano dentro do seu ambiente cultural. Foi possível cumprir dois dos três objetivos principais propostos. O único objetivo não alcançado foi identificar alterações nos dados devido a pandemia de Covid-19. Ele não foi alcançado devido à ausência de alterações relevantes nos dados para o ano de 2020. No entanto, o objetivo principal, prever o índice de felicidade através dos dados da pesquisa, e o secundário, identificar a região do mundo de um determinado país apenas através da relação das métricas com o índice, foram alcançados utilizando modelos de classificação de baixa complexidade e boa explicabilidade.

11. Links

Os arquivos, códigos, artefatos e análises mais detalhadas realizadas no projeto podem ser encontradas em https://github.com/SalatielBairros/world-happiness-report. O código do *frontend* da página web utilizada para exibir os

[19] O detalhamento das importâncias dos atributos pode ser verificado na API e visualizados na página web.

resultados pode ser encontrado em https://github.com/SalatielBairros/frontend-world-happiness-report. As publicações dos projetos podem ser encontradas em:

- Página web de visualização dos resultados: https://app-tcc-world-happines-report.azurewebsites.net/
- Documentação Swagger da API com os resultados: http://tcc-world-happines-report.azurewebsites.net/docs.

12. Referências

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