Previsão do Nível de Felicidade com Aprendizagem Supervisionada a partir de Indicadores Socioeconómicos (2020-2023)

Prediction of Happiness Level using Supervised Learning from Socioeconomic Indicators (2020-2023)

Rodriguez, 25608; Oliveira, 24845
Licenciatura em Engenharia Informática, ESTIG
Instituto Politécnico de Beja
Beja, Portugal
25608@stu.ipbeja.pt
24845@stu.ipbeja.pt

Resumo — **Este trabalho aplica técnicas de Data Mining para prever o nível de felicidade de países (Baixa/Média/Alta) a partir de indicadores socioeconómicos integrados num Data Warehouse desenvolvido previamente. O conjunto de dados final contém 548 registos ao nível país-ano (2020–2023), limpos e sem valores em falta. A variável-alvo foi construída a partir do HappinessScore (0–10) através de discretização por frequência igual (k=3). Foram treinados e comparados quatro modelos supervisionados no Orange: Decision Tree, k-Nearest Neighbors, Logistic Regression e Random Forest. A avaliação foi realizada com validação cruzada estratificada de 10 folds, recorrendo a métricas como Accuracy, Precision, Recall e F1 (macro). Os resultados mostram diferenças de desempenho entre algoritmos, sendo o modelo Random Forest o que atingiu o melhor compromisso entre Precision/Recall/F1 e menor confusão entre classes adjacentes.**

Palavras Chave - Data Mining; Classificação; Felicidade; Indicadores socioeconómicos; Orange; Validação cruzada.

Abstract — **This work applies Data Mining techniques to predict countries' happiness level (Low/Medium/High) from socioeconomic indicators integrated in a previously built Data Warehouse. The final dataset contains 548 country-year records (2020–2023), cleaned with no missing values. The target variable was derived from the 0–10 HappinessScore using equal-frequency discretization (k=3). Four supervised classifiers were trained and compared in Orange: Decision Tree, k-Nearest Neighbors, Logistic Regression, and Random Forest. Models were evaluated with stratified 10-fold cross-validation using Accuracy, Precision, Recall and (macro) F1. Results highlight performance differences across algorithms, with achieving the best balance between Precision/Recall/F1 and reduced confusion between adjacent classes.**

Keywords - Data Mining; Classification; Happiness; Socioeconomic indicators; Orange; Cross-validation.

INTRODUÇÃO

A felicidade e o bem-estar das populações têm sido objeto de crescente interesse por parte de investigadores, governos e organizações internacionais, uma vez que estão fortemente relacionados com fatores económicos, sociais e de saúde. A disponibilidade de grandes volumes de dados socioeconómicos provenientes de diferentes fontes permite a aplicação de técnicas de Data Mining e Machine Learning para a análise e compreensão dos fatores que influenciam os níveis de felicidade dos países ao longo do tempo.

Neste contexto, a aprendizagem supervisionada tem vindo a assumir um papel relevante na identificação de padrões e na previsão de fenómenos complexos, permitindo classificar observações com base em conjuntos de atributos previamente definidos. A classificação do nível de felicidade dos países constitui um problema desafiante, uma vez que envolve múltiplas variáveis interdependentes e diferentes escalas de valores, exigindo metodologias adequadas para garantir resultados fiáveis e interpretáveis.

O objetivo principal deste trabalho consiste em aplicar técnicas de aprendizagem supervisionada para classificar o nível de felicidade dos países em três categorias (Baixa, Média e Alta) com base em indicadores socioeconómicos.

METODOLOGIA

Processos (KDD)

A metodologia adotada neste trabalho segue o processo de Knowledge Discovery in Databases (KDD): (1) seleção dos dados a partir do Data Warehouse, (2) pré-processamento e validação, (3) transformação do alvo, (4) aplicação de algoritmos de classificação e (5) avaliação e interpretação.[1]

Seleção e preparação dos dados:

Para este trabalho, foi utilizado o conjunto de dados resultante do trabalho prático anterior (TP1), correspondente a uma tabela facto ao nível país-ano, já previamente integrada e limpa. O conjunto de dados final é composto por 548 instâncias, não apresentando valores em falta.

No processo de preparação, procedeu-se à definição do papel de cada variável no contexto da aprendizagem supervisionada.

As variáveis socioeconómicas foram consideradas como atributos preditores (features), incluindo indicadores económicos, sociais e de qualidade de vida, tais como rendimento, suporte social, esperança média de vida saudável, liberdade de escolha, generosidade e perceção de corrupção. As variáveis Country e Year foram mantidas como metadados (metas), uma vez que permitem a idntificação da instância (Fig1).. A variável HappinessScore, foi transformada no atributo alvo do problema originando três classes equilibrada.

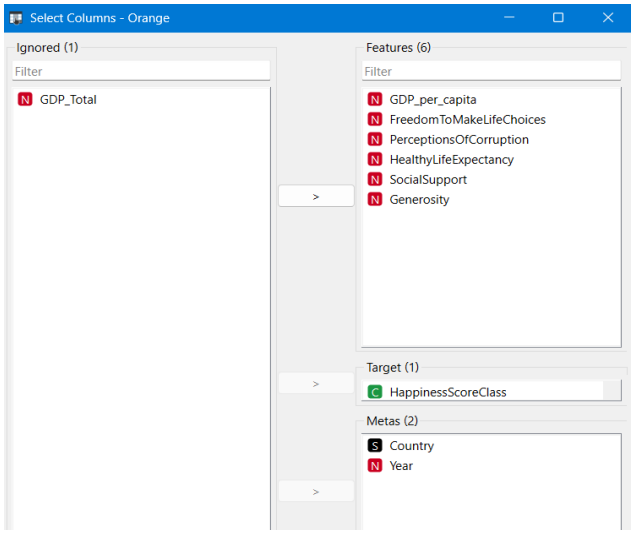

Figure 1. Definição de Roles no dataset

Todo este processo de seleção e preparação foi realizado recorrendo à ferramenta Orange

Definição do atributo alvo (3 classes)

O conjunto de dados original inclui a variável HappinessScore, medida numa escala numérica contínua, que representa o nível médio de felicidade associado a cada país num determinado ano. Para transformar o problema num cenário de aprendizagem supervisionada por classificação, foi necessário converter esta variável contínua num atributo categórico (classes).

Para esse efeito, recorreu-se ao widget Discretize do Orange, aplicando o método Equal frequency com k = 3(Fig 2), que divide a distribuição do HappinessScore em três intervalos com aproximadamente o mesmo número de instâncias. Esta opção foi escolhida para reduzir o risco de classes desbalanceadas. O processo gerou automaticamente três limites de corte, criando três categorias interpretáveis: Baixa, Média e Alta.

Após a discretização, a variável resultante foi validada/confirmada no widget Edit Domain (Fig 3), onde o HappinessScore passa a ter tipo categórico e três valores possíveis (Baixa/Média/Alta). Finalmente, no widget Select Columns (Fig 1), esta nova variável categórica foi definida como target .

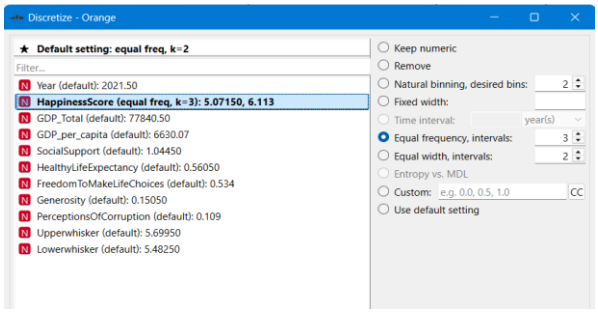

Figure 2. Discretização do HappinessScore

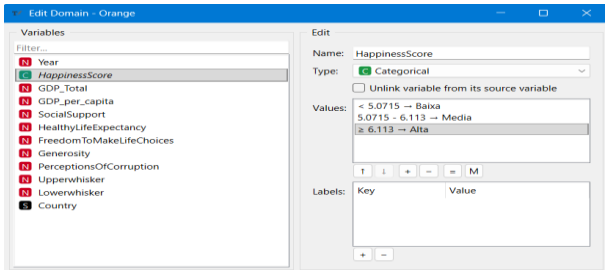

Figure 3. Conversão do atributo para categórico

Modelos e validação

Após a definição do atributo alvo, procedeu-se à aplicação de diferentes algoritmos de aprendizagem supervisionada para classificação, com o objetivo de comparar abordagens distintas e identificar o modelo com melhor desempenho na previsão do nível de felicidade dos países

Inicialmente, foi aplicada a Regressão Logística, este algoritmo permite estimar a probabilidade de pertença de cada instância às diferentes classes, foi utilizada como um modelo de referência para comparação com abordagens mais complexas.[2]

De seguida, foi considerada a Árvore de Decisão, modelo baseado em regras que permite representar o processo de decisão de forma hierárquica e intuitiva.[3]

O algoritmo k-Nearest Neighbors (kNN) foi igualmente aplicado, representando uma abordagem baseada em distância. Neste caso, a classificação de uma instância é determinada com base nas classes dos seus vizinhos mais próximos no espaço das características.[4]

Por fim, foi utilizado o algoritmo Random Forest, um método de ensemble que combina múltiplas árvores de decisão com o objetivo de melhorar a capacidade de generalização e reduzir o sobreajuste[5]

A utilização destes quatro algoritmos permitiu comparar diferentes paradigmas de aprendizagem supervisionada proporcionando uma análise mais completa e fundamentada do problema em estudo.

A validação dos modelos foi realizada utilizando a técnica de validação cruzada estratificada de 10 folds (Test & Score), assegurando uma distribuição semelhante das classes em cada fold.

Para a avaliação do desempenho dos modelos, foram consideradas várias métricas: accuracy (CA) foi utilizada como medida global da percentagem de classificações corretas métricas de precision, recall e F1-score permitiram avaliar o

desempenho dos modelos por classe, sendo o F1-score particularmente relevante por representar um compromisso entre precisão e sensibilidade.

Adicionalmente, foi utilizada a área sob a curva ROC (AUC), que mede a capacidade discriminativa dos modelos. As matrizes de confusão foram igualmente analisadas, permitindo identificar os principais tipos de erro cometidos pelos modelos e compreender quais as classes mais difíceis de prever.[6]

A combinação destas métricas e da estratégia de validação adotada permitiu uma comparação rigorosa e equilibrada entre os diferentes algoritmos testados, servindo de base para a seleção do modelo final apresentado na secção de resultados.

Resultados e Discussão

Nesta secção são apresentados os resultados obtidos pelos quatro modelos de classificação testados. A Tabela I resume o desempenho comparativo dos algoritmos, com base nas métricas selecionadas (AUC, CA, Precision, Recall, F1 e MCC). Para facilitar a leitura, dá-se especial destaque ao F1-score (média sobre classes), por refletir um compromisso entre precisão e sensibilidade, e à matriz de confusão do modelo com melhor desempenho, de modo a identificar os principais padrões de erro entre as classes Baixa, Média e Alta..

TABLE I. Resultados de Classificação(10-fold stratified cross-validation; average over classes)

Modelo	*AUC*	*CA(Accuracy)*	*F1*	*Precision*	*Recall*	*MCC*
kNN	0.837	0.743	0.734	0.737	0.743	0.619
Tree	0.889	0.754	0.757	0.764	0.754	0.632
Logistic Regression	0.908	0.766	0.769	0.777	0.766	0.652
Random Forest	0.915	0.796	0.795	0.795	0.796	0.694

O Random Forest obteve o melhor desempenho global em todas as métricas principais (CA=0.796; F1=0.795; MCC=0.694), seguido da Logistic Regression (CA=0.766; F1=0.769). A Decision Tree apresentou desempenho intermédio (CA=0.754; F1=0.757), enquanto o kNN foi o modelo com pior desempenho (CA=0.743; F1=0.734), possivelmente devido à sensibilidade do método à escala e à distribuição das features. Relativamente à capacidade discriminativa medida por AUC, observa-se novamente vantagem dos modelos Logistic Regression (0.908) e Random Forest (0.915) face a Tree (0.889) e kNN (0.837). A análise de comparação baseada na AUC (matriz de probabilidades do Orange) sugere que Logistic Regression e Random Forest apresentam probabilidade muito elevada de superar kNN e Tree. Entre Random Forest e Logistic Regression, a diferença é menor, sendo o Random Forest ligeiramente superior, mas sem uma separação tão marcada quanto nas comparações com os modelos restantes.

A. ROC Analysis (AUC)

Relativamente à capacidade discriminativa, medida por AUC, verifica-se novamente a superioridade de Random Forest e Logistic Regression face a Decision Tree e kNN. A Fig. 4 apresenta as curvas ROC (one-vs-rest) correspondentes aos modelos, confirmando visualmente a tendência observada na Tabela I: os modelos Random Forest e Logistic Regression apresentam uma separação média entre classes mais consistente.

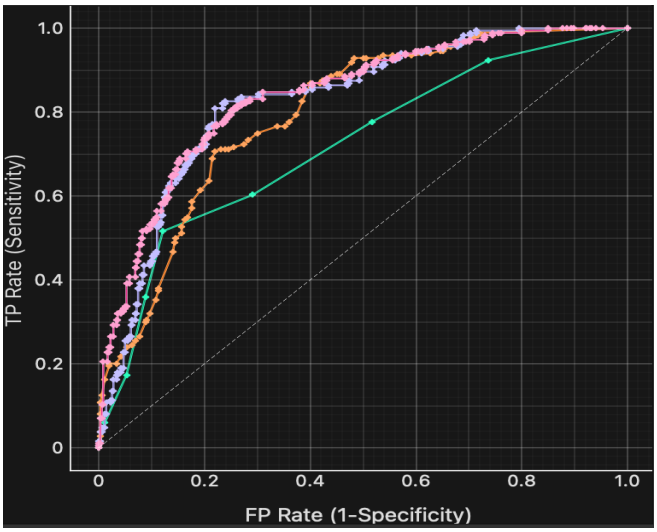

Figure 4. ROC Analysis (one-vs-rest) para os modelos avaliados; AUC médio por classes

B. *Matriz de confusão do melhor resultado*

A Fig. 5 apresenta a matriz de confusão obtida para o Random Forest. Os valores da diagonal correspondem às instâncias corretamente classificadas em cada classe (Alta/Baixa/Média). Os principais erros observados ocorreram entre classes adjacentes (Alta↔Média e Baixa↔Média), não se verificando confusões diretas entre Alta e Baixa.

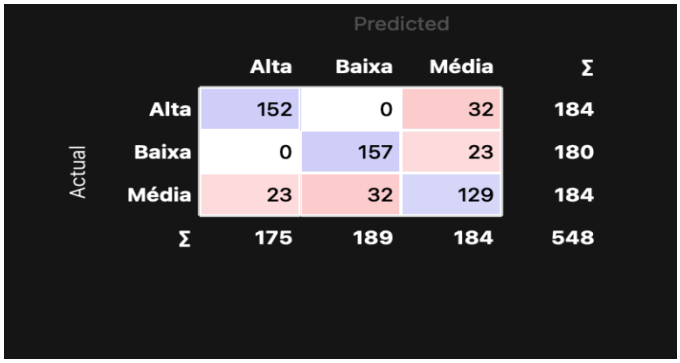

Figure 5. Matriz de confusão do Random Forest (classes: Alta/Baixa/Média)

A matriz mostra que os extremos Alta e Baixa são mais facilmente distinguíveis, não ocorrendo confusões diretas entre estas classes. Em contraste, a classe Média concentra a maioria dos erros, sendo confundida com Alta ou Baixa. Este padrão é compatível com a própria construção das classes por discretização do HappinessScore, onde a classe intermédia representa uma zona de transição e pode incluir países com perfis socioeconómicos semelhantes aos das classes adjacentes.

C. Ajuste do Random Forest

O modelo Random Forest foi configurado com 110 árvores ("Number of trees = 110"). Este valor foi definido empiricamente através de testes incrementais: inicialmente foram avaliados valores em intervalos largos (de 100 em 100 até 500) e, após identificar a zona de melhor desempenho, foram feitos refinamentos com incrementos menores (por exemplo 20 em 20) até se chegar ao valor final. Esta estratégia permitiu selecionar um parâmetro de complexidade que maximiza o desempenho sem recorrer a um espaço de procura excessivamente grande.

D. Predições do melhor modelo

Para ilustrar o comportamento do modelo final (Random Forest), a Fig. 6 apresenta exemplos de predições geradas no Orange (widget Predictions). Em cada instância, o sistema mostra as probabilidades atribuídas a cada classe na forma p(Alta):p(Baixa):p(Média) e a classe prevista (seta "→"). Observa-se que existem casos com elevada confiança, por exemplo Switzerland (2020) com 1.00:0.00:0.00 → Alta, e casos mais ambíguos em que a probabilidade é repartida por classes adjacentes (por exemplo Japan (2023) com 0.59:0.00:0.41 → Alta), o que é consistente com a maior dificuldade do problema na zona intermédia das classes.

Show probabilities for: Classes known to the model — Restore Original Order

	Random Forest	Country	Year	GDP_Total	GDP_per_capita	SocialSupport	althyLifeExpectar	lomToMakeLifeCh	Generosity	:eptionsOfCorrup
1	0.19 : 0.01 : 0.80 → Média	Japan	2021	5039148	40058.5	0.949	0.838	0.504	0.020	0.192
2	0.02 : 0.76 : 0.21 → Baixa	Cambodia	2022	39838	2325.03	0.732	0.505	0.740	0.166	0.068
3	0.59 : 0.00 : 0.41 → Alta	Japan	2023	4213167	33766.5	1.396	0.622	0.556	0.009	0.207
4	0.00 : 0.88 : 0.12 → Baixa	Namibia	2021	12399	4412.84	0.801	0.262	0.411	0.091	0.059
5	1.00 : 0.00 : 0.00 → Alta	Switzerland	2020	741288	85897.8	1.470	1.040	0.630	0.270	0.410
6	0.00 : 0.97 : 0.03 → Baixa	Zambia	2021	22096	1127.16	0.552	0.231	0.487	0.227	0.074
7	0.03 : 0.19 : 0.78 → Média	Mongolia	2023	20315	5838.63	1.494	0.244	0.425	0.239	0.058
8	0.01 : 0.95 : 0.03 → Baixa	Mauritania	2020	8464	1795.76	1.110	0.430	0.190	0.130	0.120
9	0.01 : 0.91 : 0.08 → Baixa	Jordan	2021	46362	4183.5	0.685	0.583	0.455	0.079	0.150
10	0.00 : 0.46 : 0.54 → Média	Cameroon	2021	45036	1672.33	0.556	0.159	0.425	0.205	0.058
11	0.00 : 0.92 : 0.08 → Baixa	Senegal	2020	24534	1461.09	0.950	0.520	0.350	0.160	0.080
12	0.01 : 0.05 : 0.94 → Média	Albania	2023	23388	8575.17	0.951	0.480	0.549	0.133	0.037

Figure 6. Exemplos de predições do Random Forest (probabilidades p(Alta):p(Baixa):p(Média) e classe prevista), obtidas no Orange (widget Predictions).

Como limitações do estudo, destacam-se: (i) o intervalo temporal relativamente curto (2020–2023), (ii) a transformação do HappinessScore contínuo em classes (perda de informação), e (iii) possíveis efeitos culturais/regionais não totalmente capturados pelas variáveis usadas. Como trabalho futuro, propõe-se testar normalização para kNN, aplicar seleção de features/importance para justificar variáveis mais relevantes, e comparar modelos adicionais (p.ex., Gradient Boosting), bem como analisar explicitamente o impacto do ano/região como variáveis explicativas na distinção entre classes adjacentes.

COMPARAÇÃO COM PROJETOS RELACIONADOS

De forma complementar, foi realizada uma análise comparativa entre o presente trabalho e um projeto disponibilizado na plataforma Kaggle, intitulado World Happiness Analysis and Prediction [7]. Nesse projeto, o problema é abordado como uma tarefa de aprendizagem supervisionada por regressão, tendo como objetivo a previsão direta do valor contínuo do Happiness Score, recorrendo a modelos e métricas típicas de regressão.

No nosso trabalho, optou-se por uma abordagem distinta, através da transformação do Happiness Score contínuo num conjunto de classes discretas (Baixa, Média e Alta), reformulando o problema como uma tarefa de classificação multiclasse. Esta opção metodológica permitiu analisar os níveis de felicidade de forma categorizada, facilitando a interpretação dos resultados e a aplicação de métricas específicas de classificação, como precision, recall, F1-score e AUC.

A discretização do atributo alvo implica uma alteração na forma como o desempenho dos modelos é avaliado, uma vez que o foco deixa de ser a minimização do erro numérico e passa a centrar-se na correta distinção entre categorias. Embora esta abordagem implique a perda de algum detalhe quantitativo, contribui para uma análise mais robusta face a pequenas variações no valor do indicador e favorece a identificação de padrões globais associados a diferentes níveis de felicidade.

Desta forma, as duas abordagens apresentam perspetivas complementares sobre o mesmo fenómeno, sendo a escolha entre regressão contínua e classificação discreta dependente dos objetivos do estudo e do tipo de análise pretendida.

CONCLUSÕES

Este trabalho demonstrou a aplicabilidade de técnicas de classificação supervisionada para estimar níveis de felicidade (Baixa/Média/Alta) com base em indicadores socioeconómicos. A avaliação, realizada com validação cruzada estratificada (10 folds), permitiu comparar quatro algoritmos e selecionar o modelo mais adequado. O Random Forest apresentou desempenho superior, sugerindo vantagem de métodos *ensemble* na captura de relações possivelmente não lineares entre variáveis explicativas e o nível de felicidade. A análise detalhada dos erros mostrou um padrão consistente: a classe Média é a mais difícil de classificar, concentrando confusões com Baixa e Alta, enquanto os extremos são mais separáveis. O resultado final inclui um workflow reprodutível no Orange e um conjunto de métricas e visualizações que suportam a validação do modelo. Como trabalho futuro, propõe-se reforçar a distinção da classe intermédia através de otimização de hiperparâmetros, normalização (para kNN) e avaliação de abordagens alternativas.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," in Proc. KDD, 1996.

[2] Orange Data Mining, "Logistic Regression," Orange Visual Programming Documentation.[Online].vailable:https://orange3.readthedocs.io/en/3.5.0/widgets/model/logisticregression.html (accessed: 04-Jan-2026).

[3] Orange Data Mining, "Tree — Orange Visual Programming Documentation." [Online]. Available: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/tree.html (accessed: 04-Jan-2026)

[4] Orange Data Mining, "kNN (k-Nearest Neighbors) — Orange Visual Programming Documentation." [Online]. Available: https://orange3.readthedocs.io/projects/orange-visual-

programming/en/latest/widgets/model/knn.html (accessed: 04-Jan-2026).

[5] Orange Data Mining, “Random Forest — Orange Visual Programming Documentation.”[Online].Available: https://orange3.readthedocs.io/en/3.5.0/widgets/model/randomforest.html (acesso em 04-Jan-2026).

[6] Orange Data Mining, “ROC Analysis — Orange Documentation.” [Online].Available: https://orange.readthedocs.io/en/latest/widgets/rst/evaluate/rocanalysis.html (acesso em 04-Jan-2026)

[7] DevraAI. World Happiness Analysis and Prediction. Kaggle Notebook. Disponível em: https://www.kaggle.com/code/devraai/world-happiness-analysis-and-prediction. Acedido em: janeiro de 2026.