

Relatório

Mel Yukari

June 2025

1 Introdução

Esse relatório irá apresentar os resultados encontrados por meio da regressão logística aplicada em um dataset, a fim de prever a personalidade, introvertido ou extrovertido, das pessoas. A regressão logística é uma técnica estatística usada para modelar a probabilidade de ocorrência de um evento binário, ou seja, com duas possíveis saídas, como por exemplo: sim/não, sucesso/fracasso, positivo/negativo. Um dos principais pontos fortes da regressão logística reside em sua interpretabilidade. O modelo permite compreender de forma clara como cada variável independente influencia a probabilidade do desfecho analisado. É uma técnica simples, eficiente e serve como base para muitos modelos mais complexos, como redes neurais e outras variantes da regressão.

Descrição da Base de Dados

O dataset selecionado para o trabalho foi o “Extrovert vs. Introvert Behavior Data”. Sua extração foi realizada a partir da plataforma online Kaggle, no formato CSV. Nele, encontramos informações tanto qualitativas quanto quantitativas.

A base de dados tem como intuito analisar e classificar traços de personalidade, distinguindo entre introvertidos e extrovertidos, com base em padrões comportamentais e sociais.

Dimensão dos Dados

O dataset é formado por 8 colunas (variáveis) e 2900 linhas (registros), mas para a análise, foi utilizada a variável *Going_outside*, que contém a frequência que uma pessoa sai de casa, sendo um indicador de sociabilidade. A partir dessa coluna ocorrerá a previsão de personalidade.

Metodologia

Inicialmente, foi utilizado um conjunto de dados chamado `personality_dataset.csv`, que contém informações sobre o tipo de personalidade dos participantes e sua frequência de saídas de casa. Após o carregamento do conjunto de dados, realizou-se a padronização dos nomes das colunas para o português, a fim de facilitar a interpretação dos resultados. A variável categórica “Personalidade” foi convertida em uma representação binária: indivíduos introvertidos foram codificados como 0 e extrovertidos como 1.

Para avaliar a associação entre a frequência de sair de casa e a probabilidade de um indivíduo ser extrovertido, foram aplicadas duas abordagens complementares de regressão logística:

1. **Regressão Logística com Scikit-learn:**

Os dados foram divididos em dois subconjuntos – 80% para treino e 20% para teste. Um modelo de regressão logística foi treinado com os dados de treino e, em seguida, utilizado para prever os rótulos do conjunto de teste. A performance do modelo foi avaliada por meio de uma matriz de confusão e de um relatório de classificação, contendo métricas como acurácia, precisão, revocação (recall) e F1-score.

2. **Regressão Logística com Statsmodels:**

Foi ajustado um segundo modelo utilizando a biblioteca `statsmodels`, que permite uma análise estatística mais robusta dos coeficientes da regressão. Este modelo teve como variável preditora a frequência de sair de casa e como variável dependente a codificação binária da personalidade. A partir deste ajuste, foi possível calcular as probabilidades previstas de extroversão com base nos valores da variável explicativa.

Os resultados obtidos foram representados graficamente por meio de duas principais visualizações:

- **Matriz de Confusão:** ilustrou a capacidade do modelo em classificar corretamente os indivíduos como introvertidos ou extrovertidos.
- **Curva da Regressão Logística:** apresentou a relação estimada entre a frequência de sair de casa e a probabilidade de ser extrovertido, sobreposta aos dados observados, permitindo uma visualização clara da tendência captada pelo modelo.

Resultados

Com base nos dados coletados, o modelo de regressão logística foi ajustado para estimar a probabilidade de um indivíduo ser extrovertido em função da frequência com que sai de casa. A amostra utilizada na etapa de teste foi composta por 567 indivíduos, divididos entre as categorias de personalidade introvertida e extrovertida.

	precision	recall	f1-score	support
Introverso	0.92	0.93	0.92	268
Extroverso	0.94	0.92	0.93	299
accuracy			0.93	567
macro avg	0.93	0.93	0.93	567
weighted avg	0.93	0.93	0.93	567

Figure 1: Relatório de classificação contendo precisão, recall e F1-score para cada classe.

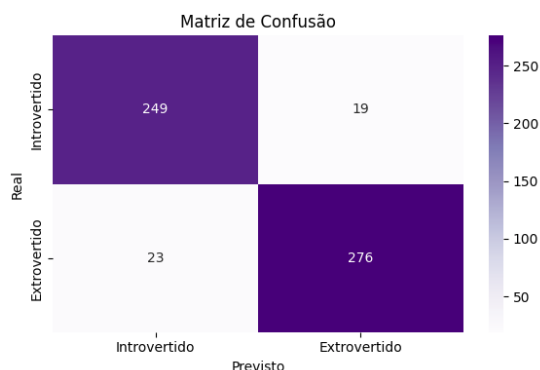


Figure 2: Matriz de confusão do modelo de regressão logística.

Relatório de Classificação

Matriz de Confusão

A matriz revelou que 249 introvertidos e 276 extrovertidos foram corretamente classificados. Ocorreram 19 erros para introvertidos e 23 para extrovertidos.

Curva da Regressão Logística

A curva mostra um padrão sigmoidal: indivíduos que saem de casa com pouca frequência (< 3 vezes/semana) tendem à introversão, enquanto os que saem mais de 6 vezes/semana são majoritariamente extrovertidos.

Conclusão

Os resultados indicam que a frequência com que um indivíduo sai de casa está significativamente associada ao seu tipo de personalidade. A regressão logística mostrou que maiores frequências de saída aumentam a probabilidade de classificação como extrovertido, em concordância com a literatura, que relaciona extroversão a maior busca por estímulos externos e interação social.

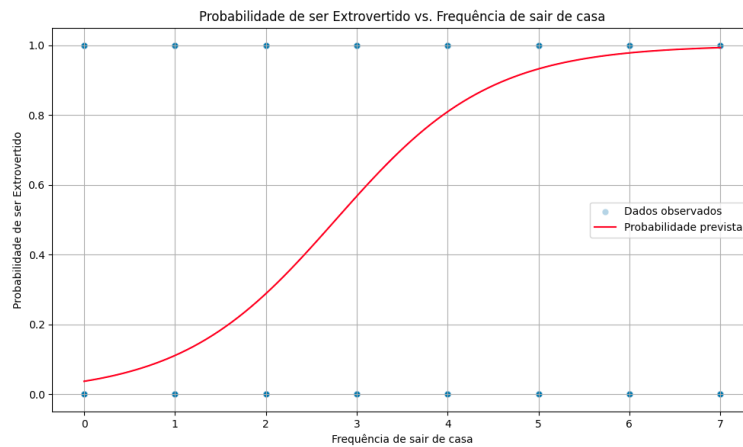


Figure 3: Probabilidade prevista de ser extrovertido conforme a frequência de sair de casa.

Apesar de se basear em uma única variável preditora, o modelo demonstrou bom desempenho, sugerindo que comportamentos observáveis podem ser úteis na inferência de traços de personalidade. No entanto, a personalidade é um fenômeno complexo, influenciado por múltiplos fatores, e o modelo não deve ser visto como instrumento diagnóstico definitivo.

Este estudo reforça o potencial de abordagens estatísticas na investigação de aspectos psicológicos, promovendo a integração entre ciência de dados e ciências humanas.

Referências

- <https://github.com/melyukari/trabalho-calculo.git>