# UNIVERSIDADE FEDERAL DO PARANÁ

EDUARDO TAKASHI HIRANO

JOÃO PEDRO ROVIDA FURTADO DE SOUSA

**ANÁLISE DAS CARACTERÍSTICAS DETERMINANTES NA PREVISÃO DAS CLASSES**:

UM ESTUDO SOBRE FATORES INFLUENCIADORES NA DISSOLUÇÃO MATRIMONIAL

2023

CURITIBA

# OBJETIVO

O objetivo primordial desta análise consiste em identificar as características mais influentes na previsão das classes e, ao mesmo tempo, investigar as possíveis correlações entre os atributos, com o propósito de detectar fatores de atenção que possam estar relacionados com o desfecho de uma dissolução matrimonial.

# METODOLOGIA

1. O Método *Gottman*

O Método *Gottman*, é, segundo o site do Instituto *Gottman*, uma abordagem à terapia de casal que inclui uma avaliação completa do relacionamento do casal e integra intervenções baseadas em pesquisas. O método aponta quatro fatores principais para a dissolução de um matrimônio:

1. Crítica à personalidade do parceiro;
2. Defensividade;
3. Obstrução ou recusa a interagir;
4. Desprezo.

Os objetivos do Método *Gottman* incluem aumentar os comportamentos de proximidade e amizade, abordar conflitos de forma produtiva e construir juntos uma vida de significado compartilhado, se fundando em sete pilares:

1. **Construir Mapas do Amor**: Envolve manter uma compreensão contínua do mundo do parceiro ao longo do tempo, incluindo seus pensamentos, sentimentos, vida cotidiana e valores, a fim de fortalecer a relação.
2. **Expressar carinho e admiração**: Casais bem-sucedidos são capazes de apreciar e desfrutar a maioria dos comportamentos de seus parceiros e aprender a conviver com suas diferenças.
3. **Voltarem-se um para o outro**: A comunicação com interesse e respeito, mesmo em tópicos triviais, é essencial para a felicidade do casal. Manter uma proporção favorável de expressar interesse e reconhecimento em comparação com ignorar estratégias de comunicação é crucial.
4. **Aceitar influência**: Levar em consideração as preferências do parceiro e estar disposto a ceder e adaptar-se ajuda a construir confiança e equilíbrio de poder no relacionamento.
5. **Resolva problemas que podem ser resolvidos**: Casais que resolvem questões usam táticas como um início suave, reparos na conexão emocional, calma, compromisso e foco no bem-estar do relacionamento.
6. **Gerenciar conflitos e superar impasses**: O método *Gottman* ajuda os casais a administrar, não necessariamente a resolver, conflitos, reconhecendo que muitos conflitos são perpétuos e fazem parte do relacionamento compartilhado.
7. **Criar significado compartilhado**: A conexão no relacionamento se fortalece à medida que ambos os parceiros enriquecem a vida um do outro, encontrando significado e dando sentido às lutas por meio de uma história compartilhada.

A Escala de Preditores de Divórcio (DPS) é uma ferramenta de avaliação psicológica desenvolvida para identificar e mensurar fatores que podem influenciar o sucesso ou o fracasso de um relacionamento conjugal.

A terapia de casais de *Gottman* é conhecida por sua ênfase na observação do comportamento dos casais em situações do dia a dia e na identificação de padrões de interação que podem ser prejudiciais ou benéficos para o relacionamento além da pesquisa extensiva sobre o casamento e o divórcio. (GOTTMAN et al., 2018)

1. Heurísticas

Neste estudo, empregamos um conjunto de técnicas de aprendizado de máquina para análise comparativa de desempenho. O foco principal recai sobre a utilização de heurísticas de árvore de decisão, a serem contrastadas com os algoritmos *Random Forest* e *k-Nearest Neighbors* (k-NN) na previsão de divórcios.

Para realização da análise exploratória, inicialmente, os dados foram importados em um ambiente *Jupyter Notebook*. Além disso, para a implementação dos modelos de árvore de decisão, *Random Forest* e k-NN, importamos as classes *Decision Tree Classifier*, *Random Forest Classifier* e *K Neighbor Classifier* da biblioteca *scikit-learn*. A avaliação dos modelos gerados foi conduzida tanto por meio da biblioteca *scikit-learn* quanto do *software Weka*, proporcionando uma abordagem abrangente para a avaliação dos resultados.

# K-NN - Histórico

O KNN (*K-nearest neighbors*), também conhecido como IBK (*Instance-Based* K), é um método clássico de aprendizado de máquina que é um algoritmo de classificação que opera com base na proximidade entre instâncias em um espaço de características. Ele é particularmente notável por sua simplicidade e eficácia em várias aplicações. Ao longo dos anos, tornou-se um dos algoritmos mais comuns na mineração de dados, destacando-se por sua facilidade de implementação e pela capacidade de obter resultados assertivos.

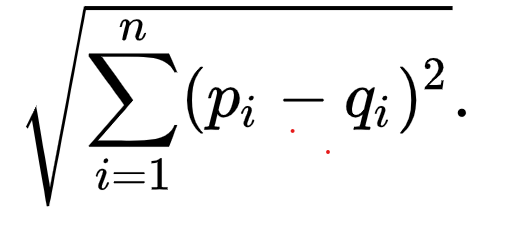
O KNN tenta classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos vizinhos mais próximos. Se os vizinhos mais próximos forem majoritariamente de uma classe, a amostra em questão será classificada nesta categoria. Diversas métricas, como a distância Euclidiana, de *Manhattan* ou *Minkowski*, são aplicadas para medir a proximidade.

Ao longo dos anos, o KNN tem sustentado sua importância em diversas áreas, incluindo reconhecimento de padrões, mineração de dados, bioinformática e processamento de imagens. Seu legado como uma abordagem intuitiva e flexível continua a moldar o desenvolvimento de novas técnicas e a compreensão do campo de *Machine Learning*.

KNN - Algoritmo

O algoritmo *K-nearest neighbors* (KNN) é um método de classificação intuitivo e eficaz que segue um procedimento passo a passo, que percorre essas 5 etapas:

1. **Recebimento do Novo Dado:** Inicialmente, o algoritmo recebe um dado não classificado que precisa ser categorizado
2. **Cálculo das Distâncias:** Em seguida, a distância entre o novo dado e cada um dos outros dados já classificados é medida. A métrica de distância escolhida, como a distância Euclidiana, de *Manhattan* ou *Minkowski*, influenciará esse cálculo. Dentre as formas mais comuns par medir distância está a Euclidiana.



1. **Seleção das Menores Distâncias:** As K menores distâncias são então selecionadas, onde K representa o número de vizinhos mais próximos a serem considerados. A escolha de K é um elemento crucial que impacta diretamente na decisão de classificação.
2. **Verificação das Classes dos Vizinhos Próximos:** Em seguida, o algoritmo verifica as classes dos dados que possuem as K menores distâncias. Isso envolve contabilizar a quantidade de vezes que cada classe apareceu entre os vizinhos mais próximos.
3. **Classificação do Novo Dado:** Com base na contagem, o novo dado é classificado como pertencente à classe que mais apareceu entre os vizinhos mais próximos.

Ao aplicar o KNN, é importante considerar a escolha da métrica de distância, a determinação do valor de K e a possibilidade de normalizar os dados para evitar influências desproporcionais antes mesmo de iniciar a metodologia. Uma boa prática para esse método é utilizar valores ímpares para K, a fim de evitar empates dentro da base de dados.

KNN - Exemplo Prático  
  
 Considere um conjunto de dados simples contendo informações sobre alturas e pesos de indivíduos, classificados como "Baixo" ou "Alto" com base em uma característica específica. Queremos classificar um novo indivíduo com altura 175 cm e peso 70 kg usando o método KNN com K=3.

| Indivíduo | Altura(cm) | Peso(Kg) | Classe |
| --- | --- | --- | --- |
| 1 | 160 | 50 | Baixo |
| 2 | 165 | 55 | Baixo |
| 3 | 170 | 60 | Baixo |
| 4 | 180 | 75 | Alto |
| 5 | 185 | 80 | Alto |
| 6 | 190 | 85 | Alto |

Antes de iniciarmos o passo a passo da metodologia temos que normalizar a tabela, para trabalharmos apenas com valores entre 0 e 1, e para isso dividimos todos os atributos de cada coluna, altura e peso, pelo atributo de maior valor na coluna selecionada.

| Indivíduo | Altura(cm) | Peso(Kg) | Classe |
| --- | --- | --- | --- |
| 1 | 0.8421 | 0.5882 | Baixo |
| 2 | 0.8684 | 0.6471 | Baixo |
| 3 | 0.8947 | 0.7059 | Baixo |
| 4 | 0.9474 | 0.8824 | Alto |
| 5 | 0.9737 | 0.9412 | Alto |
| 6 | 1 | 1 | Alto |

1. **Recebimento do Novo Dado:** Dado não classificado: Altura = 175 cm, Peso = 70 kg.

Calcule o valor normalizado para Altura(175 cm): 175/190 ≈ 0.9211.

Calcule o valor normalizado para Peso (70 kg): 70/85 ≈ 0.8235.

1. **Cálculo das Distâncias:** Calcule a distância entre o novo indivíduo e cada ponto no conjunto de dados usando a distância Euclidiana:
   1. Indivíduo 1 ≈ 0,2482
   2. Indivíduo 2 ≈ 0,1841
   3. Indivíduo 3 ≈ 0,1205
   4. Indivíduo 4 ≈ 0,0645
   5. Indivíduo 5 ≈ 0,1289
   6. Indivíduo 6 ≈ 0,1933
2. **Seleção das Menores Distâncias:** Menores Distâncias: 0.0645, 0.1205, 0.1289
3. **Verificação das Classes dos Vizinhos Próximos:** Respectivamente temos Alto, Baixo, Alto.
4. **Classificação do Novo Dado:** Com base na contagem da classe que mais aparece podemos classificar o novo dado como Alto.

Random Forest - Histórico

O modelo *Random Forest*, uma inovação marcante no campo do aprendizado de máquina, foi originalmente concebida por *Leo Breiman* e *Adele Cutler*. Esta abordagem revolucionária, marcada por sua capacidade de combinar as saídas de várias árvores de decisão para obter um resultado unificado, rapidamente se destacou como um algoritmo de referência. Desenvolvido para lidar tanto com problemas de classificação quanto de regressão, sua notável facilidade de uso e flexibilidade contribuíram significativamente para a disseminação e adoção generalizada dessa técnica ao longo da história do aprendizado de máquina.

*Leo Breiman* introduziu o método de *Bagging* (Agregação de *Bootstrap*) em 1996, onde uma amostra aleatória de dados é selecionada com substituição, permitindo a escolha de pontos de dados individuais mais de uma vez. Essas amostras são treinadas independentemente e, dependendo da tarefa, como classificação ou regressão, a média ou a maioria das previsões é usada para obter uma estimativa mais precisa. Essa abordagem é especialmente útil para reduzir a variância em conjuntos de dados ruidosos.

A *Random Forest*, uma extensão do *Bagging*, utiliza tanto o *Bagging* quanto a aleatoriedade de recursos para criar árvores de decisão não correlacionadas. A aleatoriedade de recursos, também conhecida como *Bagging* de Recursos ou "o método de subespaço aleatório", gera um subconjunto aleatório de recursos, diminuindo a correlação entre as árvores. Essa diferenciação é crucial, visto que as árvores de decisão tradicionais consideram todas as possíveis divisões de recursos, enquanto as *Random Forests* selecionam apenas um subconjunto, melhorando a generalização.

Random Forest - Algoritmo

Inicialmente, é criado um conjunto de árvores de decisão. Cada árvore é treinada de forma independente com uma amostra aleatória do conjunto de dados de treinamento. Essa amostra é selecionada com substituição, o que significa que alguns pontos de dados podem ser escolhidos mais de uma vez, enquanto outros podem não ser incluídos.

Ao crescer cada árvore, em cada nó, uma amostra de recursos é selecionada aleatoriamente. Esse processo cria diversidade entre as árvores, pois cada uma é treinada em um subconjunto diferente de características.

Após a construção das árvores, quando uma nova instância precisa ser classificada, cada árvore fornece uma previsão. No caso de classificação, a classe mais frequente entre as árvores é escolhida como a previsão final. Em problemas de regressão, a média das previsões das árvores é geralmente considerada.

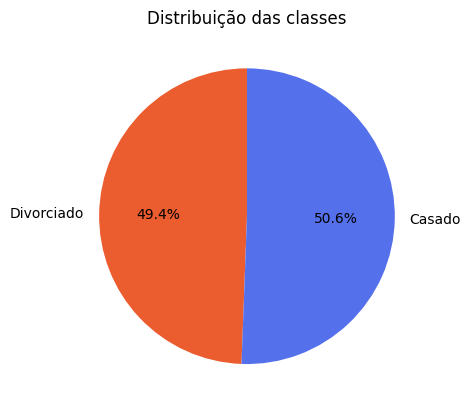
A técnica de *bagging*, juntamente com a aleatoriedade de recursos, ajuda a reduzir o *overfitting*, garantindo que as árvores não estejam fortemente correlacionadas. Isso torna o modelo robusto em diferentes conjuntos de dados e mais resistente a ruídos.

O algoritmo oferece flexibilidade em relação a hiperparâmetros, como o número de árvores na floresta e a profundidade máxima das árvores. O ajuste desses parâmetros pode ser realizado para otimizar o desempenho do modelo.

# 3. ANÁLISE EXPLORATÓRIA DOS DADOS

A base de dados recebida é referente a uma pesquisa de 54 perguntas relacionadas ao relacionamento conjugal de 170 casais com base na Escala de Preditores de Divórcio (DPS) utilizada na terapia de casais de Gottman. A amostra da pesquisa consistia em 84 (49.4%) casais que estavam divorciados e 86 (50.6%) eram casados.

**Imagem 1 - Distribuição das Classe**



Apesar do estudo citar estatísticas sobre o grupo amostral, como sexo (49% homens e 51% mulheres), faixa etária (média = 36,04, desvio padrão = 9,34) e até motivo do casamento (43,5% por amor e 56,5% por casamento arranjado), que trazem uma visão holistica da questão, o conjunto de dados em sí é composto somente pelas perguntas da DPS, onde cada resposta é avaliada em uma escala de 0 a 4, com 0 representando a avaliação mais baixa e 4 a avaliação mais alta, com o atributo meta indicando se o casal é divorciado ou não divorciado (casado). As 54 perguntas da pesquisa bem como o repositório e os *script* aplicados podem ser encontrados nos anexos.

Para uma primeira análise das características mais relevantes na predição de divorcio, foi calculado a entropia do conjunto (H = 0,9999) e o ganho de informação de cada atributo. Estão dispostas na tabela as perguntas que resultaram em um maior ganho de informação.

**Imagem 2 - Entropia dos Atributos**



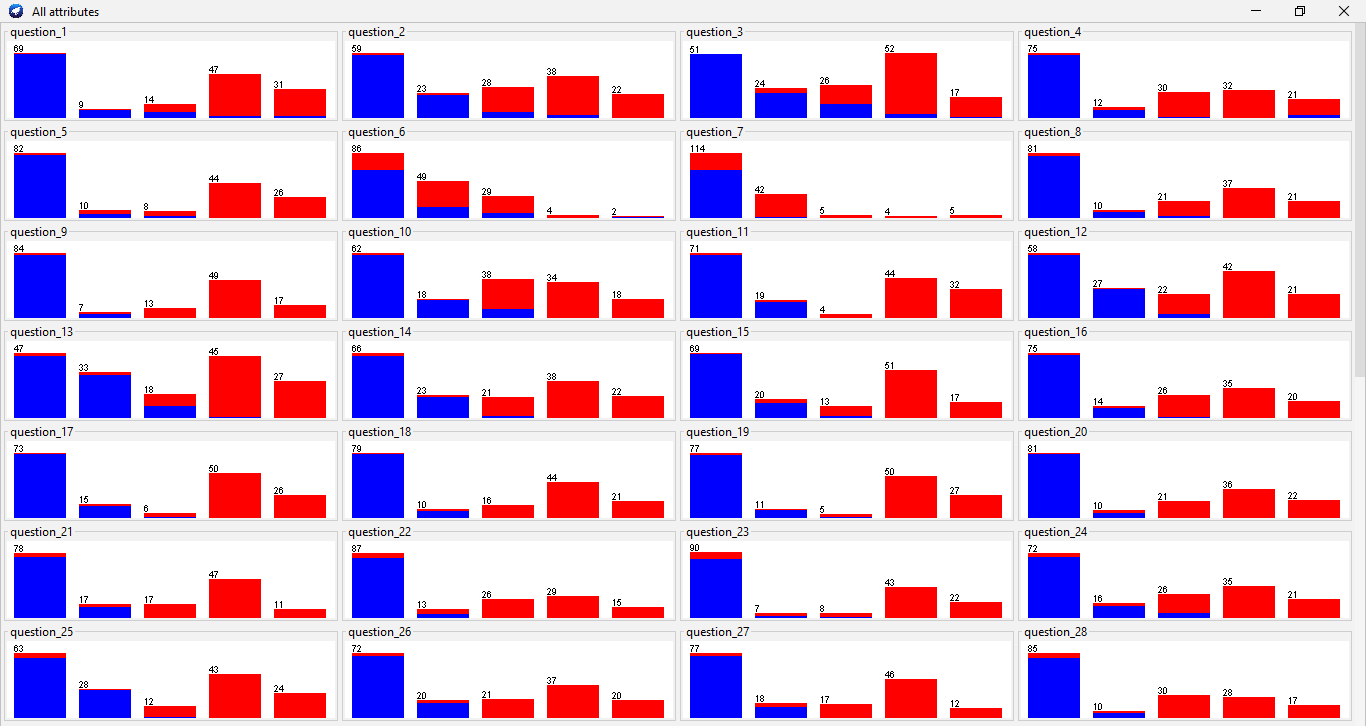
**Tabela 1 - Atributos Ranqueados por Ganho de Informação**

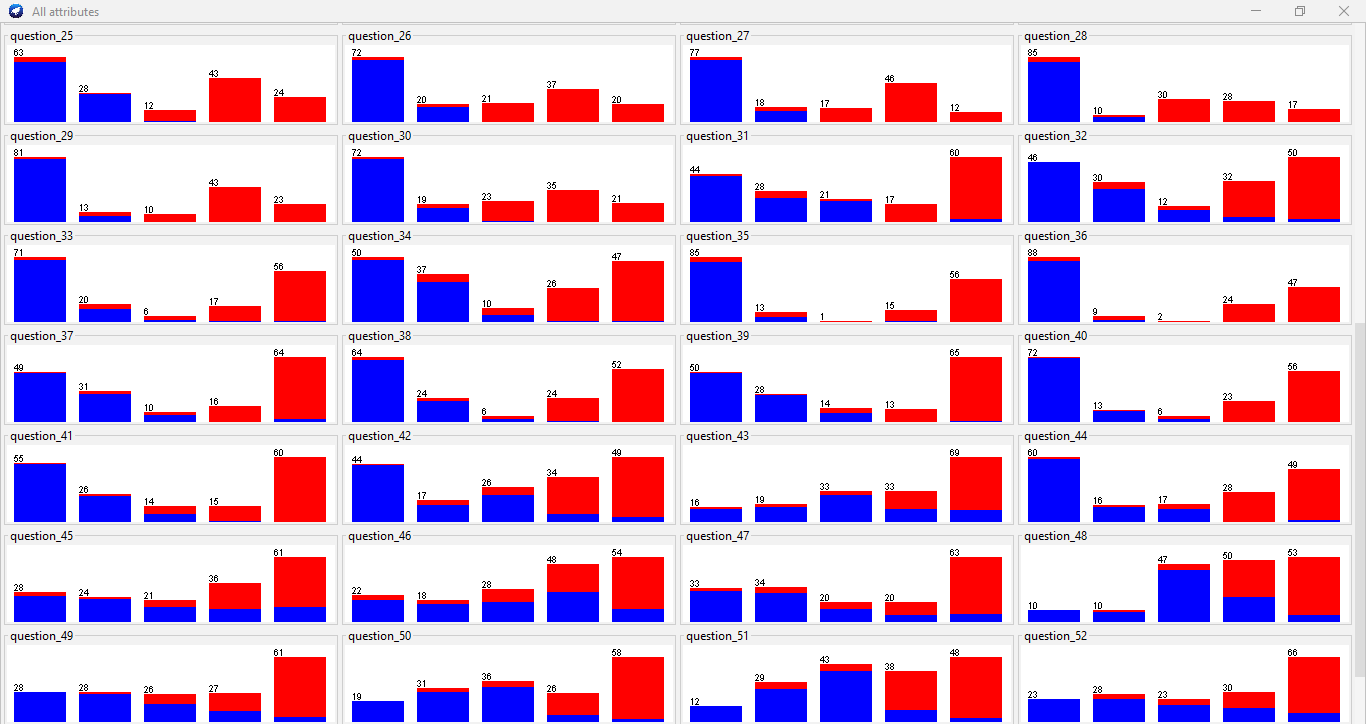
| **ranking** | **atributo** | **pergunta** | **ganho** |
| --- | --- | --- | --- |
| 1 | question 18 | Meu cônjuge e eu temos ideias semelhantes sobre como o casamento deve ser. | 0,9119 |
| 2 | question 20 | Meu cônjuge e eu temos valores semelhantes em relação à confiança. | 0,8970 |
| 3 | question 40 | Estamos apenas começando uma discussão antes que eu saiba o que está acontecendo. | 0,8900 |
| 4 | question 17 | Compartilhamos as mesmas opiniões sobre ser feliz em nossa vida com meu cônjuge. | 0,8821 |
| 5 | question 19 | Meu cônjuge e eu temos ideias semelhantes sobre como os papéis devem ser no casamento. | 0,8715 |

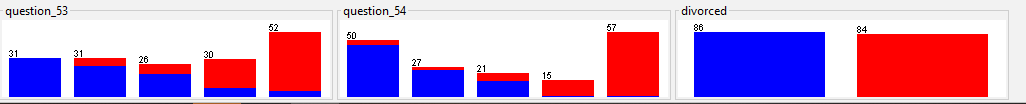
Carregando os dados no WEKA é possível começar a notar uma estranha correlação entre notas baixas e pessoas não divorciadas.

**Gráficos de Distribuição Classificada por Cor**

**(azul: Não divorciado, vermelho: Divorciado)**

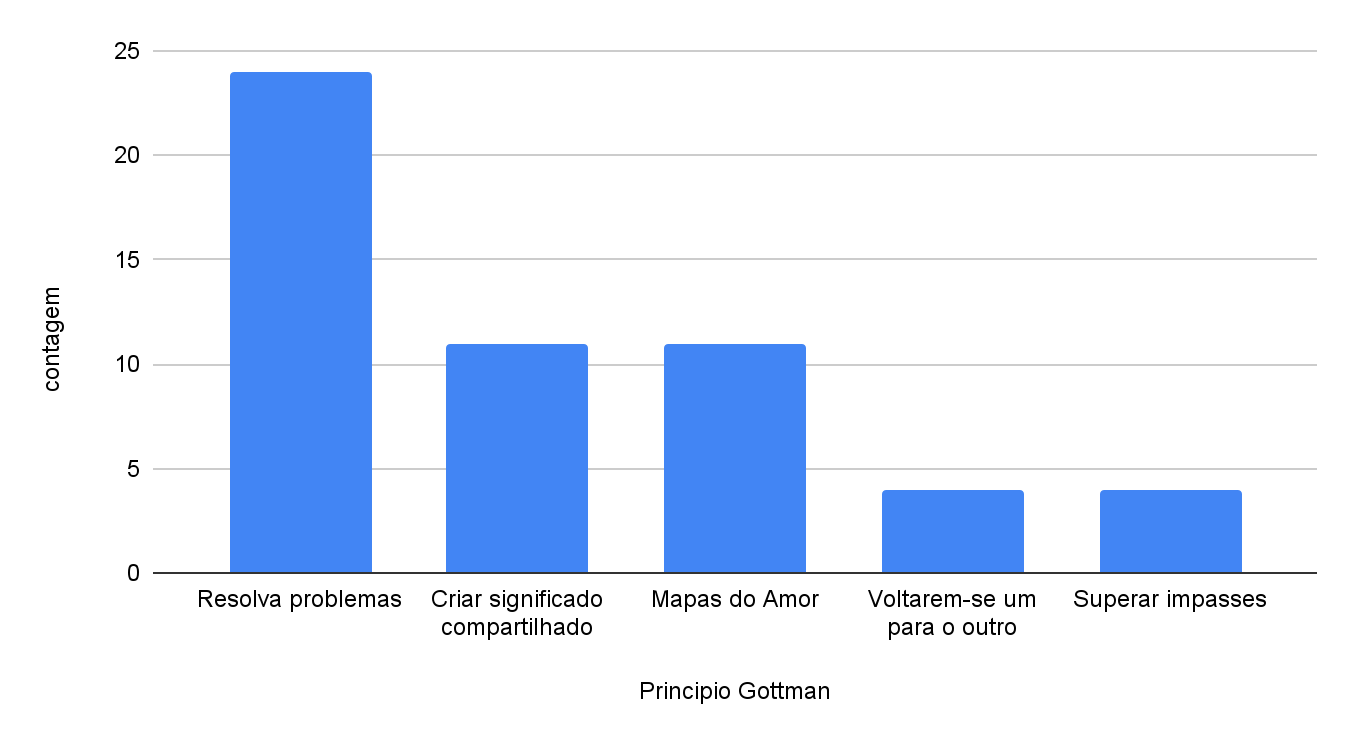




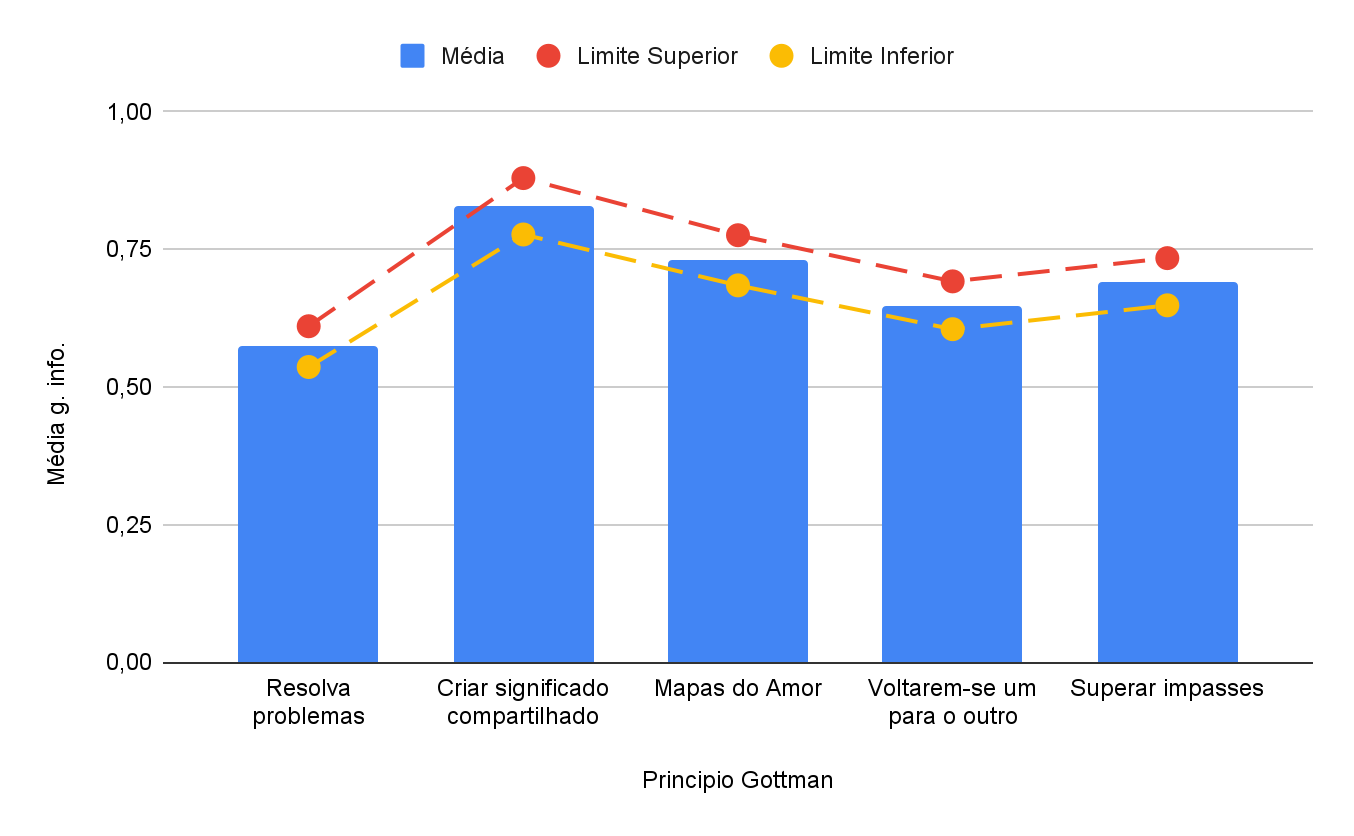


Há uma clara correlação entre as perguntas apresentadas e os sete pilares do método de Gottman. Algumas dessas perguntas podem ser identificadas no livro "Sete Princípios para o Casamento Dar Certo" (Gottman e Silver). Assim, a análise literária permitiu classificar as perguntas do questionário com base nesses princípios. Abaixo, apresento o número de ocorrências de perguntas em cada tópico, seguido pela média do ganho de informação por categoria de pergunta.

**Ocorrência de Perguntas por Categoria**



**Média de Ganho de Informação com Intervalo de Confiança de 95%**



É possível notar que, embora a categoria "Resolver Problemas" apresente a maior frequência, a média do ganho de informação nessa categoria é relativamente baixa. Até o limite superior do seu intervalo de confiança fica abaixo do limite inferior de “Criar significado compartilhado”, que emerge como a categoria mais decisiva em relação à manutenção de um casamento na população amostral.

# 4. ÁRVORE DE DECISÃO E FLORESTA ALEATÓRIA

Para efeitos de comparação, realizamos inicialmente o treinamento de um modelo de árvore de decisão com os dados. Ao analisar os resultados, observou-se uma aparente inconsistência nas informações obtidas em relação ao conjunto de dados. A expansão do caule da árvore sugere uma peculiaridade: os casais com as respostas menos favoráveis à questão de maior entropia parecem, em sua maioria, estar casados. Surpreendentemente, aqueles com as respostas mais positivas consistem predominantemente em casais divorciados. Esta discrepância parece ilógica.

**Árvore de Decisão Treinada com Conjunto de Dados com pacote Scikit-Learn do Python**

# 

|--- feature\_17 <= 1.50

| |--- feature\_25 <= 1.50

| | |--- feature\_39 <= 2.50

| | | |--- class: 0

| | |--- feature\_39 > 2.50

| | | |--- class: 1

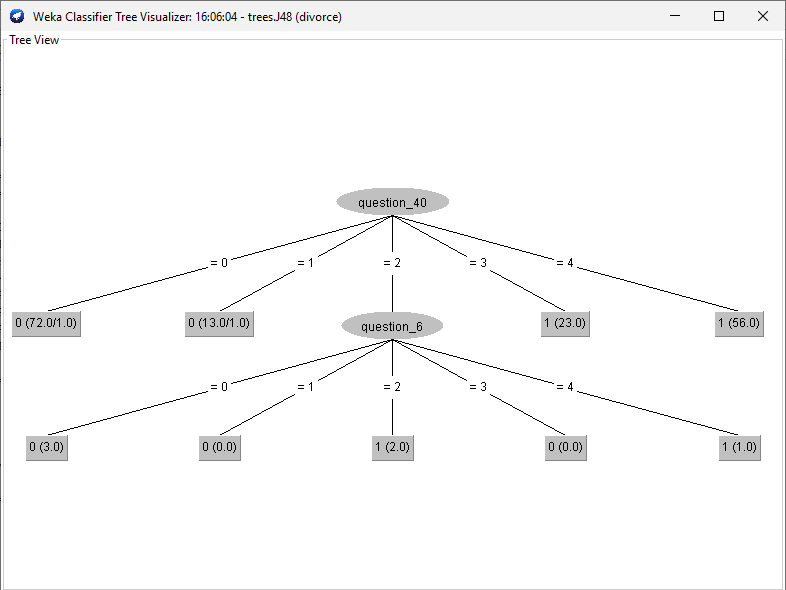
| |--- feature\_25 > 1.50

| | |--- class: 1

|--- feature\_17 > 1.50

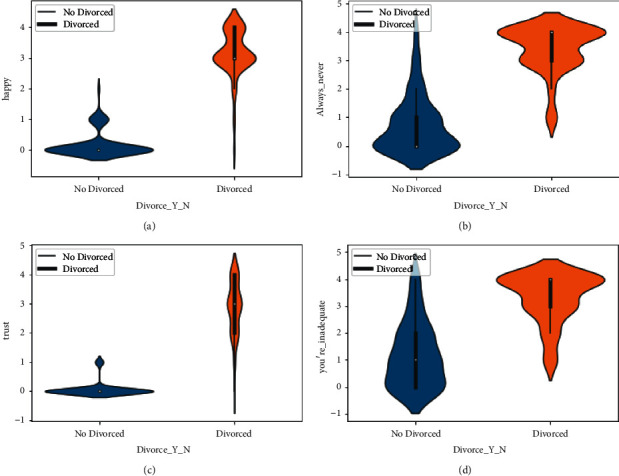
| |--- class: 1

**Árvore de Decisão Treinada com Conjunto de Dados com J48 do WEKA**



Há uma provável falta de discernimento na análise dos dados apresentada por Sadiq Fareed (2022), como evidenciado pelos gráficos de violino retirados diretamente do seu texto. Estes gráficos parecem expor abertamente as inconsistências nos dados analisados, é importante salientar que a informação de que as questões são avaliadas em uma escala de 0 a 4, onde 0 representa a pior avaliação e 4 a melhor é apresentada no artigo. Mas é surpreendente a omissão do autor perante essa incoerência.

**Gráficos de violino correspondentes às perguntas 17, 32, 20 e 53**



Abaixo estão quatro das 100 árvores que compõem a floresta aleatória.

**Imagem 6 - Amostras de Árvores da Floresta Aleatória**



É notável a melhor acurácia desse modelo perante a árvore de decisão tradicional, Enquanto a árvore de decisão atinge uma acurácia de 0,94, a floresta aleatória supera essa marca, alcançando aproximadamente 0,97.

Ao analisar a importância das características, medida pela redução de impureza, neste caso específico, pela diminuição da entropia, apresentamos a seguinte lista:

Classificação das características:

Feature 10 (0,109548) - Pergunta 11

Feature 8 (0,095410) - Pergunta 9

Feature 39 (0,093720) - Pergunta 40

Feature 19 (0,093336) - Pergunta 20

Feature 15 (0,081105) - Pergunta 16

Vale ressaltar a recorrência das perguntas 20 e 40. Contudo, é possível observar que esta avaliação proporciona uma análise mais abrangente e menos suscetível a ajustes excessivos com os dados de treinamento, abrangendo inclusive mais princípios de Gottman.

É possível constatar os parâmetros usados na criação do modelo no *script* em anexo.

# 5. VIZINHO MAIS PRÓXIMO (K-NN) E K-MEANS

Ao empregar o algoritmo KNN nos dados, obtivemos uma precisão ainda mais elevada, alcançando 0,98. Acredito que essa melhoria pode ser atribuída à aplicação do algoritmo K-Means ao modelo. Ao analisarmos os centros dos clusters gerados, percebemos dois perfis distintos: um com respostas consistentemente próximas a 0 e outro com respostas consistentemente próximas a 4.

**Coordenadas dos centros dos clusters por feature**

| features | center\_1 | center\_2 |
| --- | --- | --- |
| question 1 | 3,29 | 0,43 |
| question 2 | 2,95 | 0,50 |
| question 3 | 2,97 | 0,69 |
| question 4 | 2,82 | 0,29 |
| question 5 | 3,14 | 0,12 |
| question 6 | 1,13 | 0,41 |
| question 7 | 1,04 | 0,01 |
| question 8 | 2,93 | 0,14 |
| question 9 | 3,01 | 0,08 |
| question 10 | 2,87 | 0,42 |
| question 11 | 3,33 | 0,23 |
| question 12 | 3,05 | 0,41 |
| question 13 | 3,23 | 0,60 |
| question 14 | 2,97 | 0,32 |
| question 15 | 3,04 | 0,27 |
| question 16 | 2,94 | 0,18 |
| question 17 | 3,27 | 0,21 |
| question 18 | 3,08 | 0,13 |
| question 19 | 3,30 | 0,17 |
| question 20 | 2,99 | 0,10 |
| question 21 | 2,77 | 0,16 |
| question 22 | 2,57 | 0,07 |
| question 23 | 2,91 | 0,08 |
| question 24 | 2,91 | 0,27 |
| question 25 | 3,06 | 0,36 |

# 

# CONCLUSÃO

O conjunto de dados revelou uma clara divisão em dois grupos distintos, evidenciada pela aplicação do K-NN e pela acurácia do K-Means. Embora essa observação aponte para a possibilidade de desenvolver um modelo de avaliação robusto, é crucial destacar a presença de inconsistências nos dados em relação à definição das perguntas e à classe de predição. A lógica subjacente mostrou-se contra intuitiva, apresentando desafios na interpretação dos resultados.

Uma análise mais aprofundada e uma exploração detalhada do tema podem fornecer insights valiosos, permitindo a construção de um modelo de avaliação mais consistente. No entanto, a identificação das inconsistências nos dados ressalta a importância de uma abordagem crítica e cautelosa na interpretação dos resultados. Considerar as nuances presentes nos dados analisados é essencial para garantir a confiabilidade e a aplicabilidade das conclusões, sugerindo a necessidade de futuras investigações para esclarecer as ambiguidades observadas.

# 

# **REFERÊNCIAS**

Estatística Site (2019-09-25). Algoritmo de classificação KNN (K-Nearest Neighbors). Disponivel em: <https://estatsite.com.br/2019/09/25/algoritmo-de-classificacao-knn-k-nearest-neighbors/> Acesso em 10 de Novembro de 2023.

GOTTMAN, John M. et al. Gottman method couple therapy. Clinical handbook of couple therapy, v. 4, n. 8, p. 138-164, 2008.

GOTTMAN, John M. et al. Predicting marital happiness and stability from newlywed interactions. Journal of Marriage and the Family, p. 5-22, 1998.

GOTTMAN, John M.; SILVER, Nan. Sete princípios para o casamento dar certo. Editora Objetiva, 1999.

GOTTMAN, John M. The Gottman Method. Disponível em: https://www.gottman.com/about/the-gottman-method/. Acesso em: 10 de novembro de 2023.

IBM (2023). Florestas aleatórias. Disponível em: <https://www.ibm.com/br-pt/topics/random-forest>. Acesso em 10 de Novembro de 2023.

LAMFO/UnB (2020-07-08). Random Forest. Disponivel em: <https://lamfo-unb.github.io/2020/07/08/Random-Forest/> Acesso em 10 de novembro de 2023.

LISITSA, Elie. An Introduction to the Gottman Method of Relationship Therapy. Disponível em: https://www.gottman.com/blog/an-introduction-to-the-gottman-method-of-relationship-therapy/. Acesso em: 10 de novembro de 2023.

PINHEIRO, Juliana Baron; BOLZE, Simone Dill Azeredo. Intervenções para Terapias com Casais: Recursos Clínicos do Método Gottman. Nova Perspectiva Sistêmica, v. 30, n. 71, p. 43-55, 2021.

RIBEIRO, Alexandre P.; XAVIER, Fernando C. (2014). Ensembles de Aprendizado de Máquina. Capítulo 62. Florestas aleatórias. In: Aprendizado de Máquina: Conceitos, Algoritmos e Aplicações. 1ª ed. Porto Alegre: Bookman. p. 309-326.

SADIQ FAREED, Mian Muhammad et al. Predicting divorce prospect using ensemble learning: Support vector machine, linear model, and neural network. Computational Intelligence and Neuroscience, v. 2022, 2022.

YÖNTEM, Mustafa Kemal et al. Divorce prediction using correlation based feature selection and artificial neural networks. Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi, v. 9, n. 1, p. 259-273, 2019.

WANG, Hui. Nearest neighbors without k: a classification formalism based on probability. Faculty of Informatics, University of Ulster, 2002.