Universidade Federal de São Carlos

Centro de Ciências e de Tecnologia

Departamento de Computação

Aprendizado de Máquina 2 - Trabalho 1 - Combinação de Modelos

Professor: Murilo Coelho Naldi

Integrantes do grupo

Antonio Caique Santos, 744334, Engenharia de Computação

Jhon Wislin Ribeiro Citron, 776852, Engenharia de Computação

Lucas Henrique Marchiori, 769787, Engenharia de Computação

São Carlos, SP

2023

1. Introdução

O objetivo deste trabalho foi analisar o conjunto de dados (*Dataset*) sobre um questionário para determinar se casais se separam ou não. Além disso, aplicar o método de combinação de modelos heterogêneo e homogêneo, para observar a acurácia obtida e determinar qual modelo obteve um melhor desempenho e como a combinação de diferentes modelos de classificação afetou o resultado.

2. Conjunto de dados

O conjunto de dados escolhido pelo grupo para a análise, apresenta informações relacionadas a um conjunto de perguntas formuladas para 170 casais, para determinar se esses casais se separaram ou não. Os atributos deste *Dataset* foram basicamente as 54 questões feitas aos casais e um atributo que determinava se o casal havia se separado ou não. Outro ponto a ser comentado, é que as respostas dadas foram valores numéricos de 0 a 4 que representaram, uma escala de resposta respectivamente de Nunca (0) a Sempre (4). Esse *Dataset* pode ser consultado no link [1] das bibliografias.

3. Estudo e Tratamento do Conjunto

Todo o código aqui utilizado encontra-se disponível no *Google Colab*[1]

Ao realizar o *download* do conjunto de dados, o mesmo se encontra em dois *.csv*, um com os dados das respostas do questionário, e outro com o enunciado de cada questão feita. Por isso, é primeiramente realizada a leitura de ambos os arquivos utilizando a biblioteca Pandas[2], porém estes dois são armazenados em variáveis tipo *Dataframe* diferentes, pois não há como concatená-los em um mesmo *dataframe* sem antes processar estes dados. Após esta leitura , os conjuntos que podem ser visualizados através do comando *divorce_data.head(), reference_data.head()* estão ilustrados na figura 1 e 2 respectivamente

[1] https://colab.research.google.com/drive/1a1Axajs5Cv2Bj1DLTv9eXszJy0JCibsn?usp=sharing
[2] https://pandas.pydata.org/

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53	Q54	Divorce
0	2	2	4	1	0	0	0	0	0	0	...	2	1	3	3	3	2	3	2	1	1
1	4	4	4	4	4	0	0	4	4	4	...	2	2	3	4	4	4	4	2	2	1
2	2	2	2	2	1	3	2	1	1	2	...	3	2	3	1	1	1	2	2	2	1
3	3	2	3	2	3	3	3	3	3	3	...	2	2	3	3	3	3	2	2	2	1
4	2	2	1	1	1	1	0	0	0	0	...	2	1	2	3	2	2	2	1	0	1

Figura 1: Início do *Dataset* de respostas.

	atribute_id	description
0	1	If one of us apologizes when our discussion de...
1	2	I know we can ignore our differences, even if ...
2	3	When we need it, we can take our discussions w...
3	4	When I discuss with my spouse, to contact him ...
4	5	The time I spent with my wife is special for us.

Figura 2: Início do *Dataset* de descrição.

Além disto, outras formas de visualização que ajudaram antes da aplicação do aprendizado de máquina são realizadas através da biblioteca *ydata-profiling*[3] onde a mesma apresenta diversas maneiras de visualização de dados, e de análise exploratória de dados.

Uma das primeiras características extraídas dessa análise são as estatísticas do *dataset utilizado*, ilustrado na figura 3, algo importante a se atentar é que não existem valores faltantes, não ocorrendo necessidade de um tratamento para preencher valores vazios. Outro ponto a se observar é que todos os 55 atributos são categóricos.

[3] https://ydata-profiling.ydata.ai/docs/master/

Dataset statistics

Number of variables	55
Number of observations	170
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	15
Duplicate rows (%)	8.8%
Total size in memory	73.2 KiB
Average record size in memory	440.8 B

Variable types

Categorical	55

Figura 3: Estatísticas do Dataset.

Outra estatística apresentada é uma matriz de correlação, onde apresenta a relação entre todos os atributos do Dataset, aplicando uma escala de cores de -1 a 1. Observando a matriz e analisando a correlação das questões com o atributo divorcio, foi possível determinar que as questões Q6, Q43, Q45 e Q46 foram os atributos que possuirão uma menor correlação, em uma escala de 40% a 55%. Dessa forma, por meio da função *dirvorce_data.drop*, foi feita a remoção desses atributos do Dataset.

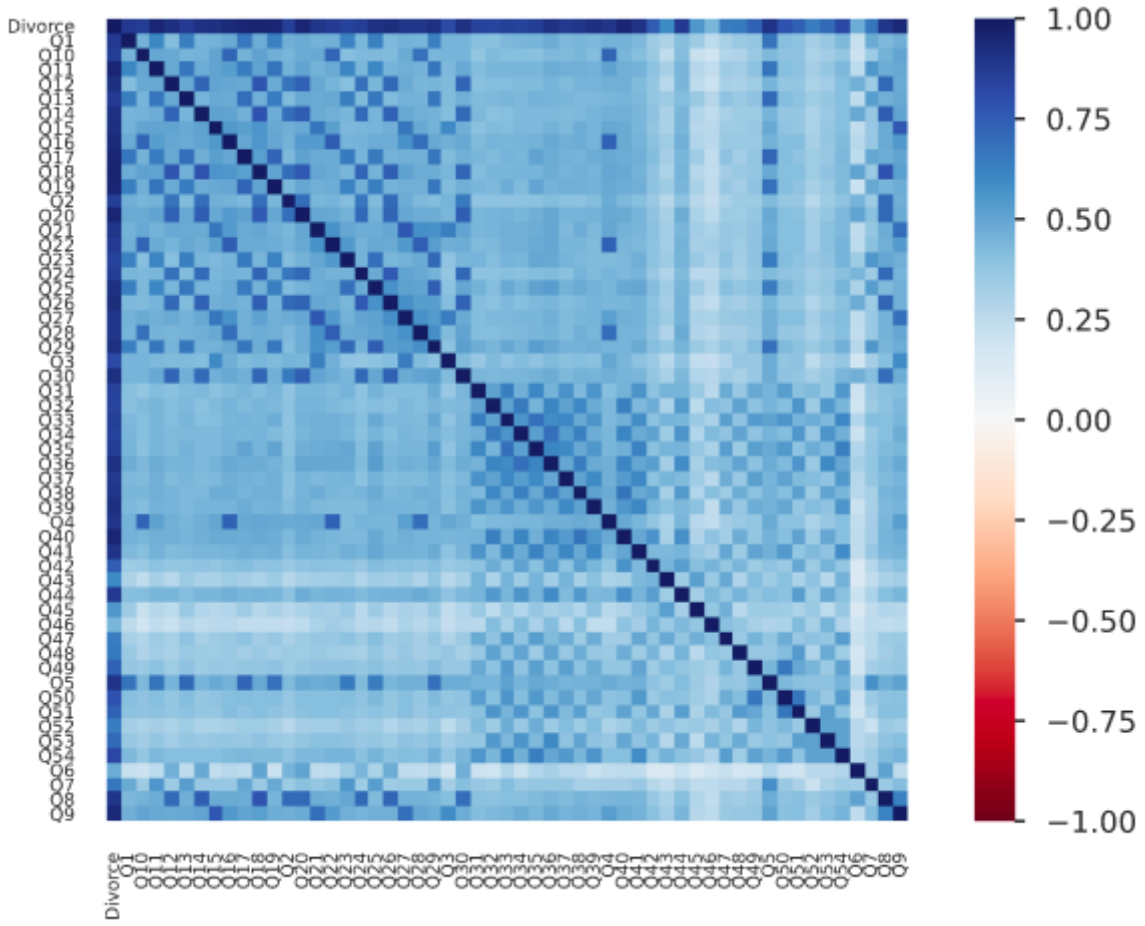

Figura 4: Heatmap de Correlação.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...
count	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	...
mean	1.776471	1.652941	1.764706	1.482353	1.541176	0.747059	0.494118	1.452941	1.458824	1.576471	...
std	1.627257	1.468654	1.415444	1.504327	1.632169	0.904046	0.898698	1.546371	1.557976	1.421529	...
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
50%	2.000000	2.000000	2.000000	1.000000	1.000000	0.000000	0.000000	1.000000	1.000000	2.000000	...
75%	3.000000	3.000000	3.000000	3.000000	3.000000	1.000000	1.000000	3.000000	3.000000	3.000000	...
max	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	...

Figura 5: divorce_data.describe()

4. Combinação Heterogênea

Feito o pré-processamento, foram criados os conjuntos de treinamento e teste para aplicar diferentes métodos de classificação (Árvore de Decisão, Vizinho mais proximo, *SVC*, *Naive Bayes Multiominal*, *MLPClassifier*), e determinar suas respectivas acurácias. A acurácia determinará o qual preciso o classificador foi. A seguir, na tabela 1, é possível observar os resultados obtidos para cada algoritmo de classificação.

Ao analisar os resultados nas Tabelas 1 e 2, é evidente que todos os algoritmos de classificação alcançaram acurácias notáveis, consistentemente superiores a 92%. Essa tendência se refletiu na combinação heterogênea de modelos, a qual alcançou uma acurácia final de 94%.

Algoritmo	**Acurácia**
Árvore de Decisão	98%
Vizinho Mais Próximo	96%
SVC	96%
Naive Bayes Multiominal	92,23%

MLPClassifier	96%

Tabela 1 - Acurácias

Modelo de Combinação	**Acurácia**
Combinação Heterogênea	94%

Tabela 2 - Combinação de Modelos Heterogênea

Em seguida, após ser feito a combinação heterogênea dos modelos, foi gerada a matriz de confusão apresentada abaixo na figura 6. A partir da matriz de confusão, foi obtido, além de uma acurácia de 96%, uma precisão de 93%, revocação de 1 e medida-f de 0,96. Vale ser ressaltado que, analisando a matriz de confusão, é possível notar que 26 instancias foram classificadas corretamente em divórcio, com 2 instancias sendo classificadas incorretamente em não divorcio, e 23 instancias classificadas corretamente em não divórcio, com 0 instancias classificadas incorretamente em divórcio.

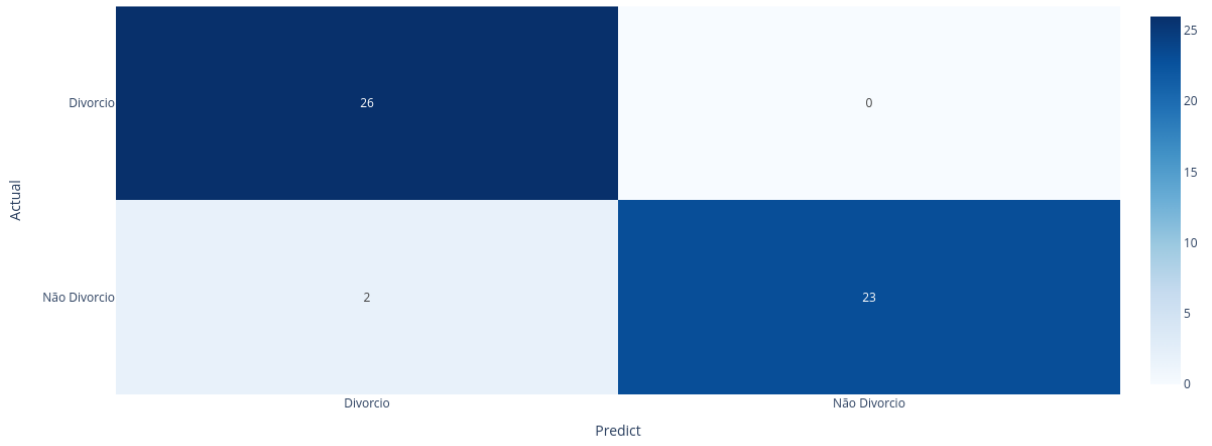

Figura 6: Matriz de Confusão (Combinação Heterogênea)

5. Combinação Homogênea

Além da implementação do modelo de combinação Heterogêneo, também foi empregado o modelo de combinação Homogêneo, permitindo uma avaliação e comparação abrangente dos resultados obtidos. Nesse contexto, foram selecionados três algoritmos de classificação. K-Nearest Neighbors (KNN), Support Vector Classifier (SVC) e Decision Tree. Para cada algoritmo, foram desenvolvidas cinco versões distintas, variando-se seus parâmetros durante o processo de modelagem. Esse enfoque sistemático possibilita uma análise mais aprofundada e robusta, visando identificar tendências e desempenhos específicos de cada abordagem.

Inicialmente, para o algoritmo Decision Tree, foram definidos para cada um dos 5 modelos criados, diferentes hiper parâmetros como a profundidade máxima (max_depth), o número mínimo de amostras necessário para dividir um nó interno (min_samples_split), o número mínimo de amostras necessário para ser uma folha (min_samples_leaf) e o número máximo de características a serem consideradas ao procurar a melhor divisão (max_features). Com isso, foram obtidos as acurácias apresentadas na tabela 3.

Ao analisar as acurácias obtidas, observamos que mesmo com diferentes configurações de hiper parâmetros, todos os modelos apresentaram elevadas taxas de acerto, consistentemente superiores a 94%. Dessa forma, ao combinar esses modelos, alcançamos uma pontuação agregada de 94%.

Decision Tree	Acurácia
dtc1	0,92
dtc2	0,94
dtc3	0,94
dtc4	0,98
dtc5	0,94
Combinação Homogênea	0,94

Tabela 3 - Decision Tree Acurácias (Homogênea)

Posteriormente, após ser feito a combinação homogênea dos modelos, foi gerada a matriz de confusão apresentada abaixo na figura 7. A partir da matriz de confusão, foi obtido, além de uma acurácia de 94%, uma precisão de 90%, revocação de 1 e medida-f de 0,95. Vale ser ressaltado que, analisando a matriz de confusão, é possível notar que 26 instancias foram classificadas corretamente em divórcio, com 3 instancias sendo classificadas incorretamente em não divorcio, e 22 instancias classificadas corretamente em não divórcio, com 0 instancias classificadas incorretamente em divórcio.

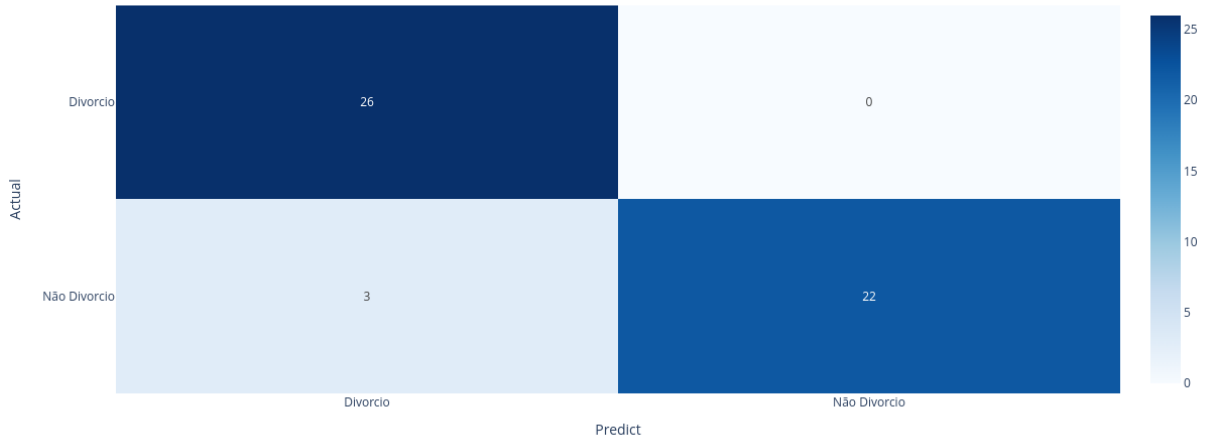

Figura 7: Matriz de confusão (Combinação Homogênea - Decision Tree)

Após realizar a combinação homogênea com a Decision Tree, procedemos de maneira análoga para o algoritmo KNN. O parâmetro n_neighbors, que determina o número de vizinhos mais próximos, foi configurado de diferentes formas. Notavelmente, a acurácia para todas as configurações de classificação permaneceu consistente, atingindo 96%. Consequentemente, a combinação homogênea manteve essa mesma acurácia.

KNN	**Acurácia**
Knn1	0,96
Knn2	0,96
Knn3	0,96
Knn4	0,96
Knn5	0,96

Combinação Homogênea	0,96

Tabela 4 - Knn Acurácias (Homogênea)

Por conseguinte, após ser feito a combinação homogênea dos modelos para KNN, foi gerada a matriz de confusão apresentada abaixo na figura 8. A partir da matriz de confusão, foi obtido, além de uma acurácia de 96%, uma precisão de 93%, revocação de 1 e medida-f de 0,96. Vale ser ressaltado que, analisando a matriz de confusão, é possível notar que 26 instancias foram classificadas corretamente em divórcio, com 2 instancias sendo classificadas incorretamente em não divorcio, e 23 instancias classificadas corretamente em não divórcio, com 0 instancias classificadas incorretamente em divórcio.

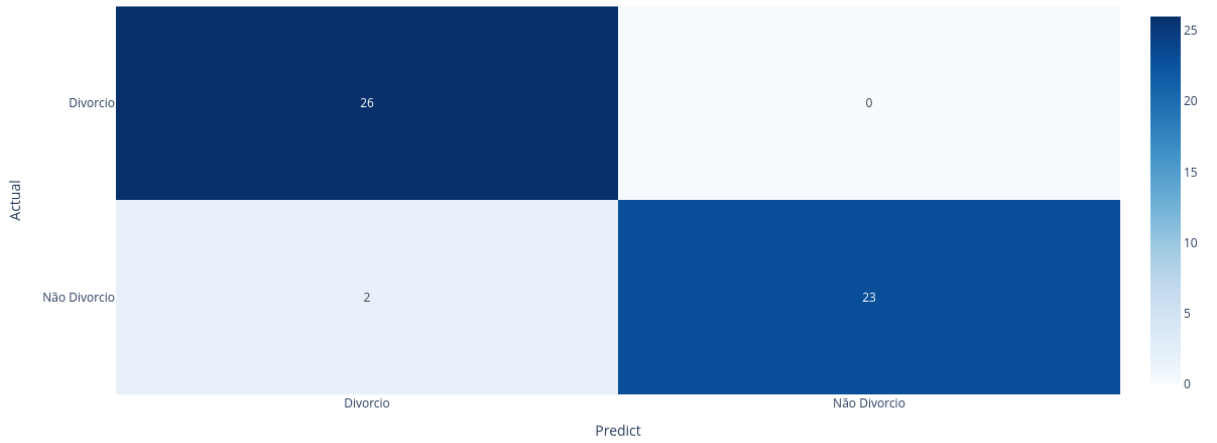

Figura 8: Matriz de confusão (Combinação Homogênea - Knn)

Após realizar a combinação homogênea com a Decision Tree e KNN, adotamos uma abordagem semelhante para o algoritmo SVC. O parâmetro kernel, que define o tipo de função kernel a ser utilizada, foi configurado de diferentes maneiras. Em resultado, três das cinco configurações predefinidas alcançaram uma acurácia de 96%, enquanto a quarta configuração obteve 45%, e a quinta, 28%. Ao ser feita a combinação homogênea dessas configurações de SVC, foi obtido uma acurácia final de 96%.

SVC	**Acurácia**
svc1	0,96
svc2	0,96

svc3	0,96
svc4	0,45
svc5	0,28
Combinação Homogênea	0,96

Tabela 5 - SVC Acurácias (Homogênea)

Por fim, após ser feito a combinação homogênea dos modelos para SVC, foi gerada a matriz de confusão apresentada abaixo na figura 9. A partir da matriz de confusão, foi obtido, além de uma acurácia de 96%, uma precisão de 93%, revocação de 1 e medida-f de 0,96. Vale ser ressaltado que, analisando a matriz de confusão, é possível notar que 26 instancias foram classificadas corretamente em divórcio, com 2 instancias sendo classificadas incorretamente em não divorcio, e 23 instancias classificadas corretamente em não divórcio, com 0 instancias classificadas incorretamente em divórcio.

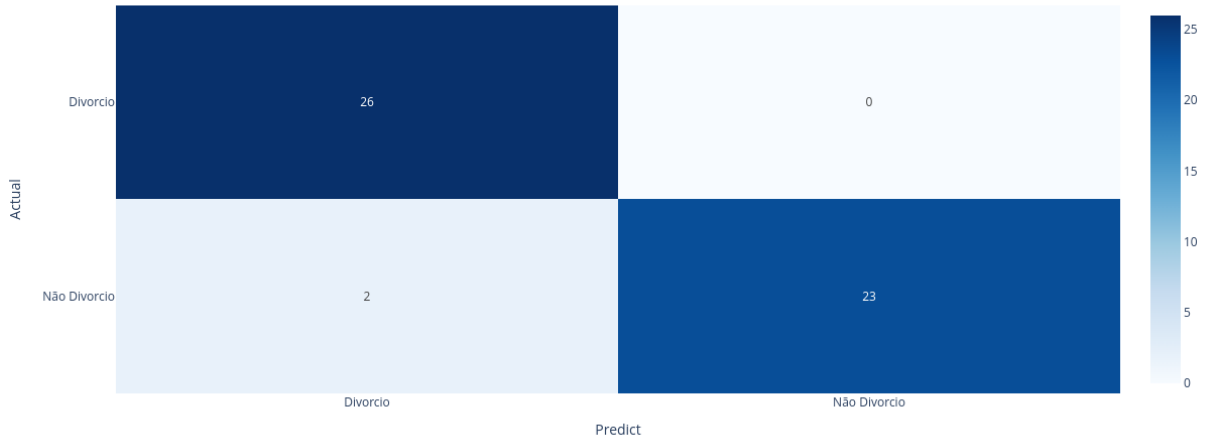

Figura 9: Matriz de confusão (Combinação Homogênea - SVC)

6. Conclusão

Após a aplicação tanto de modelos de combinação homogêneos quanto heterogêneos no conjunto de dados em questão, concluímos que, para o dataset utilizado, a estratégia de combinação de modelos não se mostrou vantajosa. As classificações individuais, sem a aplicação da combinação, já apresentaram pontuações significativamente elevadas. Nesse contexto, a combinação de modelos

não proporcionou benefícios ou melhorias nas pontuações, permanecendo essencialmente inalterada.

7. Bibliografia

[1] https://www.kaggle.com/datasets/andrewmvd/divorce-prediction