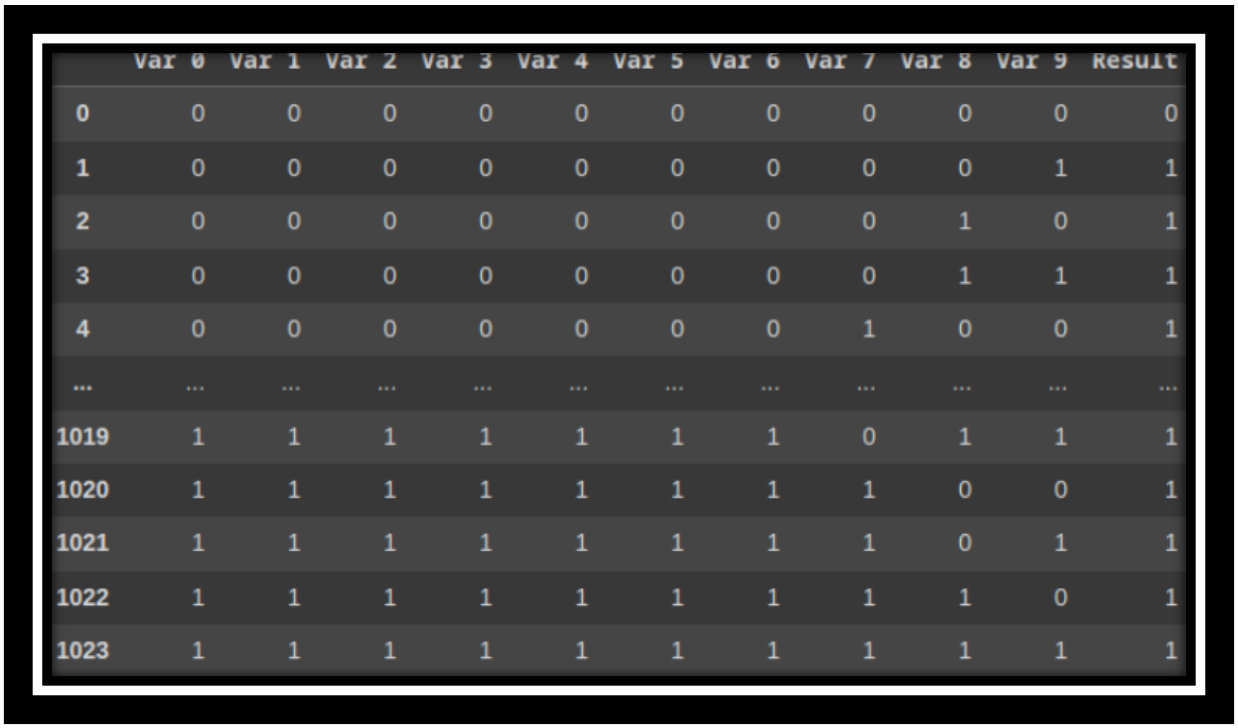
1)

Na questão, um perceptron foi usado no exercício do Colab. Ele calcula o resultado final de uma entrada usando a soma dos produtos dos valores de entrada e os pesos, incluindo o peso do viés (bias). A função de ativação é a função de limiar, que atribui 1 para sinais positivos e 0 para negativos. O erro é medido pela função de perda 0-1, que dá 0 se não há erro, e 1 se há. A saída do neurônio é dada pela fórmula do bias somado ao produto dos valores de entrada pelos seus pesos. O perceptron pode achar planos de divisão linear para classificar dados linearmente. Ele prevê corretamente funções lógicas AND e OR, mas falha com a XOR, que não é linearmente separável.

AND

	Var 0	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Result
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	1	1	0
4	0	0	0	0	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
1019	1	1	1	1	1	1	1	0	1	1	0
1020	1	1	1	1	1	1	1	1	0	0	0
1021	1	1	1	1	1	1	1	1	0	1	0
1022	1	1	1	1	1	1	1	1	1	0	0
1023	1	1	1	1	1	1	1	1	1	1	1

OR



	Var 0	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Result
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	1
2	0	0	0	0	0	0	0	0	1	0	1
3	0	0	0	0	0	0	0	0	1	1	1
4	0	0	0	0	0	0	0	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...
1019	1	1	1	1	1	1	1	0	1	1	1
1020	1	1	1	1	1	1	1	1	0	0	1
1021	1	1	1	1	1	1	1	1	0	1	1
1022	1	1	1	1	1	1	1	1	1	0	1
1023	1	1	1	1	1	1	1	1	1	1	1

XOR

	Var 0	Var 1	Result
0	0	0	1
1	0	1	1
2	1	0	0
3	1	1	0

As tabelas verdade demonstram isso:

Para OR (A || B):

Se A e B são verdadeiros, o resultado é verdadeiro.

Se A é verdadeiro e B é falso, o resultado é verdadeiro.

Se A é falso e B é verdadeiro, o resultado é verdadeiro.

Se A e B são falsos, o resultado é falso.

Código disponível em :
https://colab.research.google.com/drive/1QzWKJrX5WRwhC1gj5Jyhv2YGlkFMu82o?usp=sharing

2) Letra C

3) Letra C

4)

$0.5 * 1 + 0.4 * 1 + (-0.3 * 1) = 1$

$0.5 * 1 + 0.4 * 0 + (-0.3 * 1) = 1$

$0.5 * 0 + 0.4 * 1 + (-0.3 * 1) = 1$

$0.5 * 0 + 0.4 * 0 + (-0.3 * 1) = 0$

Letra C

5) Veja em:
https://colab.research.google.com/drive/1QzWKJrX5WRwhC1gj5Jyhv2YGlkFMu82o?usp=sharing

Pre processamento:

No pré-processamento, os atributos nominais foram convertidos em numéricos para uso em algoritmos de aprendizado de máquina, utilizando técnicas como codificação one-hot. Outliers, ou pontos de dados anômalos, foram identificados e tratados, mas não havia nenhum significativo. A normalização, que ajusta os valores para uma escala comum, não foi necessária devido à natureza categórica dos dados. O balanceamento de classes não foi mencionado como um problema, sugerindo uma distribuição equitativa de classes. Mais detalhes estão no Colab da Lista 7.

Escolha de parâmetros: Durante a avaliação e escolha dos hiperparâmetros, foram empregados métodos como Grid Search, CVParameterSelection e MultiSearch para determinar os mais eficazes. As heurísticas para definir o número de neurônios na camada oculta incluíram a regra da média, regra da raiz quadrada, regra de Kolmogorov e um valor padrão da biblioteca sci-kit learn. A taxa de aprendizado também foi determinada por esses métodos, optando-se entre três valores específicos. Detalhes adicionais estão disponíveis no Colab da Lista 7.

Resultados:

```
0.7272727272727273
```

	precision	recall	f1-score	support
0	0.82	0.79	0.81	39
1	0.53	0.56	0.55	16
accuracy			0.73	55
macro avg	0.67	0.68	0.68	55
weighted avg	0.73	0.73	0.73	55

6) O artigo mencionado oferece uma revisão detalhada sobre estratégias para tornar os modelos de aprendizado de máquina, especialmente aqueles considerados "caixas pretas" como as redes neurais profundas, mais interpretáveis. A preocupação central é que, apesar do alto desempenho, a falta de transparência nessas técnicas pode ser um impedimento, especialmente em setores onde entender as decisões tomadas pela máquina é crucial.

O texto se aprofunda em classificar e discutir as várias abordagens desenvolvidas para explicar e interpretar esses modelos. Os métodos são agrupados em:

Pós-processamento: Técnicas aplicadas após o treinamento do modelo para elucidar suas decisões.

Intrínsecos: Abordagens onde a interpretabilidade é incorporada diretamente na arquitetura do modelo.

Aproximação: Métodos que criam modelos mais simples e compreensíveis que se aproximam do comportamento do modelo original.

Cada uma dessas categorias é examinada em detalhes, considerando suas vantagens e limitações e as situações em que são mais aplicáveis. O artigo se concentra particularmente na explicação de previsões individuais, ou seja, fornecer uma justificativa compreensível para a saída de um modelo em resposta a uma entrada específica, em vez de explicar seu funcionamento interno completo.

Além de descrever essas técnicas, o artigo aborda o desafio de inspecionar e compreender o funcionamento interno dos modelos de caixa preta e as razões por trás de suas previsões. Isso envolve criar representações, visuais ou textuais, que possam elucidar o processo de tomada de decisão do modelo.

As Figuras 8 e 9 no artigo exemplificam algumas das técnicas discutidas para resolver esses problemas, enquanto a avaliação das explicações geradas também é considerada importante pelos autores. Eles discutem métricas de avaliação que podem ser usadas para medir a eficácia e a qualidade das explicações fornecidas pelos modelos.

O artigo termina com uma discussão sobre futuras tendências e desafios na área de interpretabilidade de modelos de aprendizado de máquina, enfatizando que a área está em constante evolução e que a interpretabilidade é cada vez mais demandada para aplicações críticas. O autor reforça a necessidade de pesquisa contínua para desenvolver técnicas que

possam explicar efetivamente os modelos de aprendizado de máquina, mantendo-os ao tempo precisos e confiáveis.

7) O documentário "Coded Bias" da Netflix explora os temas de viés e injustiça nos algoritmos de aprendizado de máquina. Ele destaca como os sistemas de inteligência artificial (IA), incluindo o reconhecimento facial, podem perpetuar preconceitos raciais e de gênero. O documentário segue pesquisadores como Joy Buolamwini, cujo trabalho revelou alta taxa de erros em softwares de reconhecimento facial, especialmente em rostos de mulheres negras. Isso levanta questões sobre a confiabilidade e ética do uso de IA na vida cotidiana e em sistemas de tomada de decisão críticos, como contratação de empregos, empréstimos e aplicação da lei.

A relação desse documentário com o artigo "A Survey of Methods for Explaining Black Box Models" é bastante direta. O artigo discute métodos para tornar os modelos de aprendizado de máquina mais transparentes e interpretáveis, o que pode ajudar a identificar e mitigar o viés. Ao melhorar a interpretabilidade dos modelos, os pesquisadores e desenvolvedores podem entender melhor como e por que certas decisões são tomadas, o que é crucial para assegurar que os sistemas de IA sejam justos e não discriminatórios. Ambos destacam a necessidade de abordar as implicações éticas da IA e de desenvolver tecnologia de maneira responsável.