UNIVERSIDADE FEDERAL DO PIAUÍ – UFPI
Centro de Ciências da Natureza - CCN
Programa de Pós-Graduação em Ciência da Computação - PPGCC
Disciplina: Aprendizagem de Máquina
Professor: Vinicius Ponte Machado

Relatório Trabalho Prático 1 - Árvore de Decisão

Descrição: Aplicar o Algoritmo de Árvore de Decisão
Base sorteada: (Senso) Census Income (ics.uci.edu)
Grupo 2: Danilo e Marcelo

Objetivos

Este estudo, objeto de trabalho prático da disciplina de Aprendizagem de Máquina, tem como propósito analisar o conjunto de dados "Adult" disponível no Repositório de Aprendizado de Máquina da UCI. Pretendemos desenvolver um modelo de classificação para categorizar os participantes da pesquisa com base em sua renda, utilizando variáveis características e aplicando a técnica de Árvore de Decisão. O foco é identificar as características mais relevantes na determinação se a renda anual ultrapassa ou não os $50.000 Dólares.

1. Preparação Do Ambiente

O ambiente foi montado a partir do Colab, plataforma colaborativa baseada em nuvem que simplifica o acesso e manipulação de dados, para a análise algumas bibliotecas essenciais, como pandas, numpy, matplotlib e seaborn.

2. O Dataset

O conjunto de dados "Adult" compreende informações relevantes acerca de renda, educação, idade, sexo, raça, entre outros atributos. Com 48.842 entradas distribuídas em 15 colunas, o conjunto representa diversas características individuais, conforme detalhado na Tabela 1.

A divisão prévia dos dados resultou em dois conjuntos distintos: o conjunto de treino, denominado adult.data, com 32.560 entradas, e o conjunto de testes, adult.test, com 16.282 entradas. A coleta desses dados refere-se ao ano de 1994, e a variável-alvo, Renda Anual (*income*), apresenta uma divisão entre as classes de "Alta Renda" (>50K) e "Baixa Renda" (≤50K).

Tabela 1 - Representação do conjunto de dados (Dataset) e seus respectivos atributos.

Coluna	Variável	Definição	Tipo de Variável	Tipo de Dado
age	Idade	Idade do respondente	Discreta (de 17 a 90)	int64
workclass	Classe de trabalho	Classificação do trabalho do respondente	Categórica (9 categorias)	object
fnlwgt	Peso final	O número de pessoas que acreditam no senso	Discreta	int64
education	Escolaridade	O nível de escolaridade mais elevado obtido pelo respondente	Ordinal (16 categorias)	object
education-num	Número de escolaridade	O número de escolaridade associada ao 'education' do respondente	Discreta (de 1 a 16)	int64
marital-status	Estado Civil	O estado civil do respondente	Categórica (7 categorias)	object
occupation	Ocupação	Qual o tipo de trabalho do respondente	Categórica (15 categorias)	object
relationship	Relacionamento familiar	Tipo de relacionamento familiar do respondente	Categórica (6 categorias)	object
race	Raça	A raça do respondente	Categórica (5 categorias)	object
sex	Sexo	O sexo do do respondente	Categórica (2 categorias)	object
capital-gain	Ganho de capital	Valor de ganhos de capital que o respondente obteve sobre sua poupança, investimentos e pensão	Contínua	int64
capital-loss	Perda de capital	Valor de perda de capital que o respondente obteve sobre sua poupança, investimentos e pensão	Contínua	int64
hours-per-week	Horas por semana	Quantidade de horas trabalhadas por semana pelo respondente	Discreta (de 1 a 99)	int64
native-country	País de origem	Nacionalidade do respondente	Categórica (42 países)	object
income	Renda	Classificação se o respondente é de baixa ou de alta renda	Booleano (≤ USD 50 mil, > USD 50 mil)	object

Fonte: Elaborado pelos autores (2023).

Logo no primeiro momento da avaliação do *Dataset*, verificamos a distribuição da variável alvo: a Renda. Para isto, foi criada uma função para avaliar a simetria entre os valores de cada categoria de renda.

Após o procedimento, verificamos um notável desequilíbrio na distribuição de renda entre os participantes da pesquisa. Os respondentes classificados como "Baixa Renda" (≤50K) excede em mais de três vezes o número de participantes enquadrados na categoria de "Alta Renda" (>50K).

Seguindo, aproveitamos para verificar a existência de dados faltantes, embora não tenha localizado valores faltantes, algumas categorias apresentavam dados não respondidos, a exemplo, o caractere interrogação (?). Assim, realizamos a substituição dos valores faltantes por NaN (*Not a Number*) e em seguida realizamos a contagem de valores desconhecidos há em cada coluna.

Após a contagem, e demonstrado na Tabela 2, foi identificado valores faltantes apenas nas variáveis: *workclass*, *occupation* e *native-country*, onde foi possível notar que o número de valores faltantes nas variáveis *workclass* e *occupation* são praticamente iguais. Inferindo-se que, quase sempre que houver um valor faltante em *workclass*, haverá um valor faltante em *occupation*.

Tabela 2 - Valores faltantes após contagem, com respectivo percentual.

	Valores Faltantes	Porcentagem (%)
occupation	1843	5.7
workclass	1836	5.6
native-country	583	1.8
age	0	0.0
fnlwgt	0	0.0
education	0	0.0
education-num	0	0.0
marital-status	0	0.0
relationship	0	0.0
race	0	0.0
sex	0	0.0
capital-gain	0	0.0
capital-loss	0	0.0
hours-per-week	0	0.0
income	0	0.0

Fonte: Elaborado pelos autores (2023).

3. Explorando as variáveis

Separando as variáveis numéricas e categóricas, em conjuntos de dados diferentes para tratá-las separadamente, pôde ser observado que a maioria dos respondentes são estadunidenses (29.169), isso nos levou a considerar os demais países como 'Outros'.

3.1 Explorando as variáveis numéricas

Utilizamos o método ***.describe*** do Pandas para elaborar um resumo contendo as estatísticas descritivas fundamentais de todas as variáveis numéricas. Este sumário compreende informações como os valores mínimos, médios, máximos, desvio padrão, entre outros parâmetros estatísticos relevantes, demonstrados na Tabela 3.

Tabela 3 - Sumário com parâmetros estatísticos considerados como relevantes.[1]

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32560.000	32560.000	32560.000	32560.000	32560.000	32560.000
mean	38.582	189781.814	10.081	1077.615	87.307	40.437
std	13.641	105549.765	2.573	7385.403	402.966	12.348
min	17.000	12285.000	1.000	0.000	0.000	1.000
25%	28.000	117831.500	9.000	0.000	0.000	40.000
50%	37.000	178363.000	10.000	0.000	0.000	40.000
75%	48.000	237054.500	12.000	0.000	0.000	45.000
max	90.000	1484705.000	16.000	99999.000	4356.000	99.000

Fonte: Elaborado pelos autores (2023).

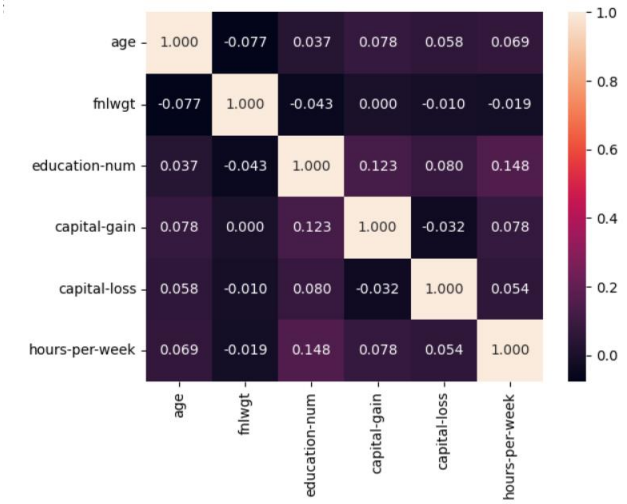
Visando proporcionar uma compreensão mais aprofundada dos dados, criamos uma representação gráfica por histogramas, para ilustrar as distribuições das variáveis numéricas e a partir dessa análise, extraímos itens a considerar, tais como:

- A distribuição da variável idade atinge seu pico para idades inferiores a 50 anos.
- O valor máximo da variável *fnlwgts* está abaixo de 400.000.
- Observa-se uma proporção bastante reduzida de indivíduos investindo em capital, com alguns casos excepcionais, como aqueles que atingem ganhos de capital superiores a US$ 90.000, entretanto, para aqueles que registraram perdas de capital, a média dessas perdas gira em torno de US$ 2.000.
- Notamos que a maioria dos participantes não apresenta perdas de capital.
- A média das horas semanais de trabalho gira em torno de 40 horas, sendo que existem outliers, incluindo casos de pessoas que trabalham quase 100 horas ou apenas 1 hora por semana.

3.2 Correlações entre as variáveis numéricas

Procedemos com a análise da correlação entre as variáveis numéricas, podendo ser observado na Figura 1, para isso, criamos uma função destinada a gerar um mapa de calor (*heatmap*) com as correlações dessas variáveis.

Figura 1 - Mapa de calor



Fonte: Elaborado pelos autores (2023).

[1] Os dados da coluna '*hours-per-week*' possuem valores mínimo de 1h e máximo de 99h. Em nossa percepção, provavelmente foram informados incorretamente ou erro na entrada dos dados.

3.3 Explorando as variáveis categóricas

Continuamos aprofundando a compreensão por meio de representação gráfica e plotamos gráficos de barra de todas as variáveis categóricas e obtivemos informações relevantes, dentre elas:

- A maioria de empregos no setor privado é evidente, com uma distribuição relativamente uniforme entre diversas ocupações governamentais e autônomas.
- A maioria dos participantes possui formação acadêmica, abrangendo diplomas de Ensino Superior, Ensino Médio ou algum nível de faculdade. Essa distribuição assemelha-se de maneira notável aos anos de escolaridade (*education-number*), sugerindo uma exploração mais aprofundada nesse ponto.
- Em relação ao estado civil, a maioria são casadas, enquanto a parcela que passou por divórcio ou viuvez é significativamente menor. Outros participantes não têm histórico de casamento.
- A distribuição das ocupações apresenta uma tendência mais uniforme, dificultando a identificação de padrões significativos, dado o número considerável de categorias.
- A maioria dos participantes assume o papel de marido ou não possui uma família constituída. O grupo étnico predominante é o branco, com os afrodescendentes representando o único grupo étnico com uma presença considerável na amostra.
- A presença majoritária de homens na amostra é demonstrada pela observação de que a maioria dos participantes são casados, evidenciando uma possível correlação entre os dois fatores.
- A grande maioria dos entrevistados tem nacionalidade estadunidense. Como mencionamos anteriormente, a disparidade entre os grupos de Baixa Renda (inferior a 50 mil) e Alta Renda (superior a 50 mil) é notória, indicando um desequilíbrio significativo nos dados.

4. Processamento dos dados

Para o pré-processamento dos dados, foram executadas as seguintes etapas:

- Remoção de Outliers: Outliers na variável de ganho de capital foram eliminados com o intuito de generalizar de forma mais eficaz os dados, promovendo maior consistência nas análises.

- Tratamento de Valores Ausentes: Para lidar com valores nulos, optou-se por remover registros que apresentavam quaisquer atributos ausentes. Essa abordagem foi adotada considerando que a exclusão de registros nulos, representando aproximadamente 5% do conjunto de dados de treinamento, não teria um impacto significativo no processo de treinamento.

- Remoção de Duplicatas: Duplicatas foram removidas para mitigar potenciais problemas de overfitting, garantindo uma representação mais fiel dos dados.

- Descarte da Variável *fnlwigt*: A variável *fnlwgt* foi descartada por ser considerada um recurso inútil, não contribuindo de forma substancial para a análise.

- Seleção de Variáveis Educação e Número de Educação: Entre as variáveis Educação e Número de Educação, optou-se por utilizar apenas "número de educação", uma vez que contém todas as informações presentes em Educação, simplificando o conjunto de dados sem perda de informação.

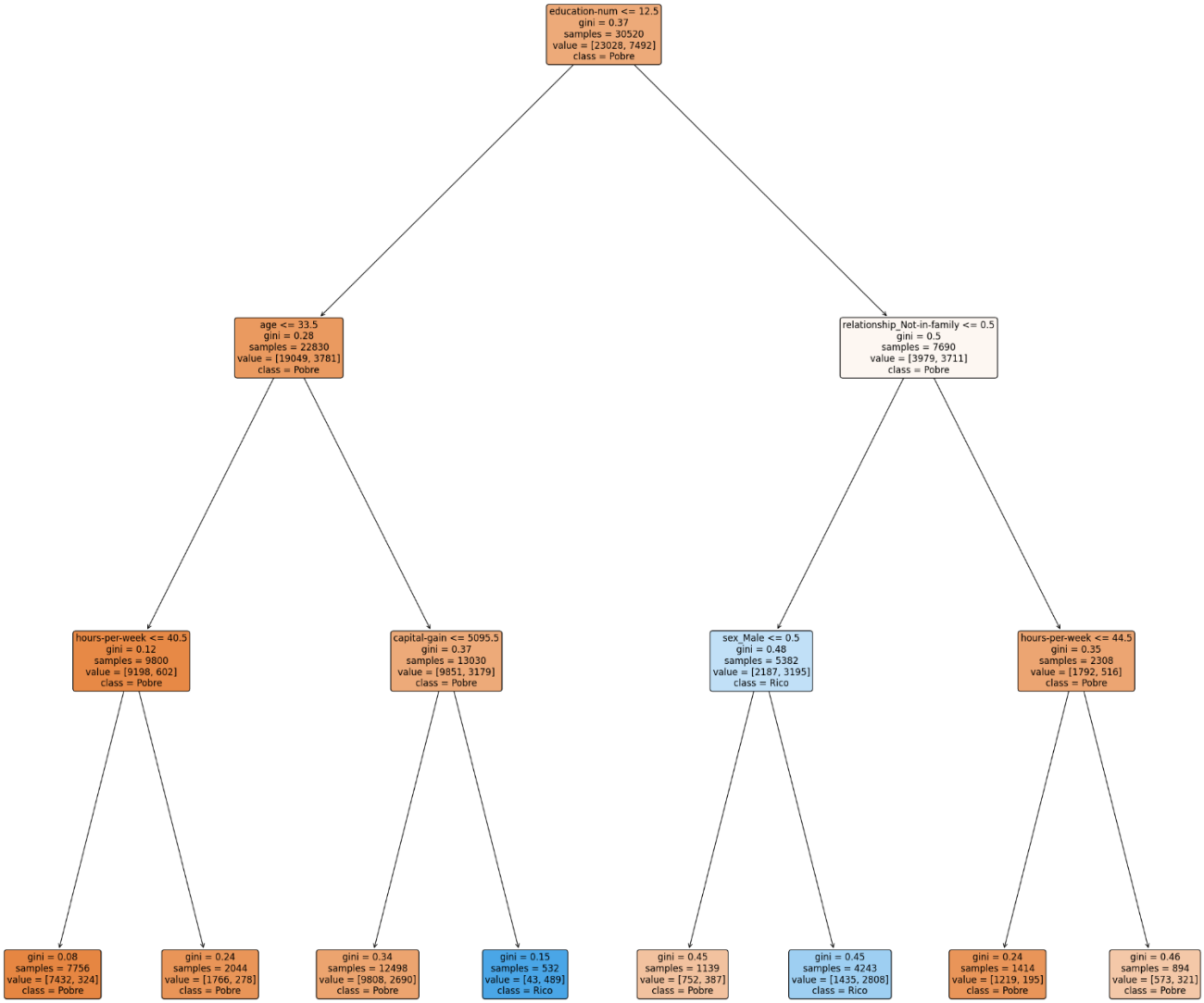
- Combinação de Ganhos e Perdas de Capital: Ganhos e perdas de capital foram combinados para formar um único recurso, facilitando a interpretação e análise conjunta desses aspectos financeiros.

- Conservação dos Demais Recursos: Todos os outros recursos foram mantidos sem alterações, preservando a integridade das informações originais.

5. Construção do modelo

Após gerado os *datasets* de treino e normalizando os dados, foi importado o modelo de Classificador por Árvore de Decisões, ilustrado pela Figura 2.

Figura 2 - Árvore de decisões



Fonte: Elaborado pelos autores (2023).

6. Conclusão

A partir da análise dos resultados da classificação, observamos que o modelo apresentou uma acurácia geral de 81%, indicando sua capacidade de predição correta para essa tarefa específica. No entanto, ao considerar métricas mais detalhadas, como *precision*, *recall e f1-score*, verificados na Figura 3, percebemos algumas disparidades.

Figura 3 - Resultados gerados (*precision*, *recall e f1-score).*

Conjunto de Teste:

	precision	recall	f1-score	support
0	0.84	0.94	0.88	11355
1	0.69	0.44	0.53	3700
accuracy			0.81	15055
macro avg	0.77	0.69	0.71	15055
weighted avg	0.80	0.81	0.80	15055

Fonte: Elaborado pelos autores (2023).

Para a classe 0 "Pobre" (renda inferior a $50.000), o modelo demonstra um grande desempenho, temos um ótimo *precision* (84%) e *recall* (94%), indicando que a maioria das previsões positivas para essa classe é assertiva, e o modelo consegue identificar grande parte dos casos verdadeiros dessa categoria.

No entanto, para a classe 1 "Rico" (renda superior a $50.000), o desempenho é inferior, com *precision* de 69% e *recall* de 44%. Isso sugere que o modelo tem mais dificuldade em identificar corretamente os casos reais de renda superior, e as previsões positivas para essa classe têm uma precisão um pouco inferior.

A análise da "*macro avg*" e "*weighted avg*" mostra que o desempenho do modelo, a partir da precisão e recordação do teste, apontando ser equilibrado quando consideramos ambas as classes, observados a partir de f1-score cujo valor foi de 80%.

Em resumo, embora o modelo tenha uma acurácia geral satisfatória, a atenção deve ser direcionada para melhorar o desempenho na identificação de casos de renda superior, buscando um equilíbrio mais robusto entre *precision* e recall para ambas as classes.