UNIVERSIDADE FEEVALE

Nome: Luisa Christ Hugendobler

Disciplina: Inteligência Artificial e Computacional

Data: 27/11/2025

1 O dataset

O repositório utilizado está localizado no link: LINK

O repositório possui dois datasets. O primeiro, possui dados de tamanho de raças de cachorro do American Kennel Club (AKC). O segundo, dog_intelligence, possui dados de uma pesquisa feita pelo Stanley Coren, professor da Universidade de Columbia. Com esses daod, é possível explorar relações entre raças de cachorro, tamanho das raças e capacidade de obediência e inteligência.

O dataset inclui as informações de raça (Breed), classificação, obediência (a probabilidade de obediência ao primeiro comando), repetições de limites (para o entendimento de comandos).

Com esses dados, será explorado a possibilidade de identificar a relação entre porte de cachorros e seu potencial de obediência. Para isso, o atributo class/target escolhido é o “obey”.

2 Caracterização dos dados

Dataset American Kennel Club (AKC)

Nome	**Tipo**	**Quantidade**	**Intervalo**
Breed	Object	150	-
height_low_inches	Object	150	6-48
height_high_inches	Object	150	8-66
Weight_low_lbs	Object	150	2-175
Weight_high_lbs	Object	150	5-190

Valores faltantes: As colunas *height_low_inches, height_high_inches, weight_low_lbs, weight_high_lbs* possuem todas 2 valores faltando.

Atributo mais desbalanceado: Não há atributo desbalanceado nesse dataset.

Quantidade de instâncias: 150 instâncias

Head do dataset:

	index	Breed	height_low_inches	height_high_inches	weight_low_lbs	weight_high_lbs
0	0	Akita	26.0	28.0	80.0	120.0
1	1	Anatolian Sheepdog	27.0	29.0	100.0	150.0
2	2	Bernese Mountain Dog	23.0	27.0	85.0	110.0
3	3	Bloodhound	24.0	26.0	80.0	120.0
4	4	Borzoi	26.0	28.0	70.0	100.0

Dataset dog_intelligence

Nome	**Tipo**	**Quantidade**	**Intervalo**
Breed	Object	136	-
Classification	Object	136	'Brightest Dogs', 'Excellent Working Dogs', 'Above Average Working Dogs', 'Average Working/Obedience Intelligence', 'Fair Working/Obedience Intelligence', 'Lowest Degree of Working/Obedience Intelligence '
Obey	Object	125	30-95
Reps_lower	Int64	136	1-81
Reps_upper	Int64	136	4-100

Valores faltantes: A coluna *obey* possui 11 valores faltando.

Atributo mais desbalanceado: O atributo Classification apresenta desequilíbrio entre suas categorias. A categoria mais frequente (“Average Working…”) possui 42 ocorrências, enquanto a menos frequente (“Brightest Dogs”) possui 10. Isso demonstra um desbalanceamento moderado.

Quantidade de instâncias: 136 instâncias

Head do dataset:

	index	Breed	Classification	obey	reps_lower	reps_upper
0	0	Border Collie	Brightest Dogs	95.0	1	4
1	1	Poodle	Brightest Dogs	95.0	1	4
2	2	German Shepherd	Brightest Dogs	95.0	1	4
3	3	Golden Retriever	Brightest Dogs	95.0	1	4
4	4	Doberman Pinscher	Brightest Dogs	95.0	1	4

3. Descrição dos procedimentos de pré-processamento

Dataset American Kennel Club (AKC)

Nesse dataset todas as colunas estavam com o tipo “Object” (string). Foi necessário converter as colunas *height_low_inches, height_high_inches, weight_low_lbs, weight_high_lbs* para numérico.

Foram criadas as colunas derivadas *height_avg* e *weight_avg*, contendo a média de altura e de peso, respectivamente.

Dataset dog_intelligence

Nesse dataset a coluna *obey* precisava ser numérica para trabalharmos com ela, porém ela era do tipo “Object” (string) contendo o símbolo de %. Foi necessário remover o símbolo % e, após, converter para numérico.

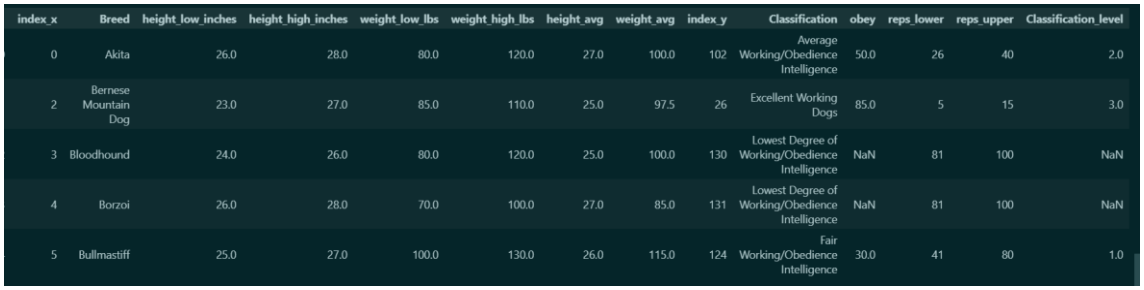
Foi criada a coluna *Classification_level* que agrupa os valores da coluna *Classification* em 3 grupos:

Categorias	**Grupo**
Brightest Dogs Excellent Working Dogs	3
Above Average Working Dogs Average Working/Obedience Intelligence	2
Fair Working/Obedience Intelligence Lowest Degree of Working/Obedicence Intelligence	1

Merge dos datasets

Após o tratamento dos dados, foi feito um merge entre os dois datasets utilizando um *inner join* com a chave Breed, coluna em comum nas duas tabelas.

Head do novo dataset:



index_x	Breed	height_low_inches	height_high_inches	weight_low_lbs	weight_high_lbs	height_avg	weight_avg	index_y	Classification	obey	reps_lower	reps_upper	Classification_level
0	Akita	26.0	28.0	80.0	120.0	27.0	100.0	102	Average Working/Obedience Intelligence	50.0	26	40	2.0
2	Bernese Mountain Dog	23.0	27.0	85.0	110.0	25.0	97.5	26	Excellent Working Dogs	85.0	5	15	3.0
3	Bloodhound	24.0	26.0	80.0	120.0	25.0	100.0	130	Lowest Degree of Working/Obedience Intelligence	NaN	81	100	NaN
4	Borzoi	26.0	28.0	70.0	100.0	27.0	85.0	131	Lowest Degree of Working/Obedience Intelligence	NaN	81	100	NaN
5	Bullmastiff	25.0	27.0	100.0	130.0	26.0	115.0	124	Fair Working/Obedience Intelligence	30.0	41	80	1.0

4. Análise de algortimos

4.1 Algoritmo Árvore de Decisão

O primeiro algoritmo escolhido foi o Árvore de Decisão. Foi utilizado duas versões do class/target: a original, numérica (de 30-95) e a categorizada (“baixa”, “média”, “baixa”). Essa variação foi escolhida devido o baixo desempenho da original.

4.1.1 Versão original

Tempo de treino	Acurácia	Precisão	Recall	F1-score	Tempo de predição
0.0039 segundos	0.2413	0.2403	0.2413	0.2397	0.0040 segundos

4.1.2 Versão categorizada

Resultados obtidos após o pré-processamento transformando a coluna “obey” numérica em categorizada seguindo a seguinte função:

```
def categorizar_obediencia(valor):
    if valor <= 50:
        return "baixa"
    elif valor <= 80:
```

```
    return "media"
else:
    return "alta"
```

Tempo de treino	Acurácia	Precisão	Recall	F1-score	Tempo de predição
0.0108 segundos	0.4137	0.3192	0.4137	0.3563	0.0020 segundos

4.2 Algoritmo KNN – K-nearest neighbors

O segundo algoritmo escolhido foi o KNN – K-nearest neighbors, que analisa os “vizinhos” mais próximos do elemento que queremos predizer. Para esse algoritmo, também foi utilizado a variação categorizada da class/target *obey*.

Como pré-processamento necessario para esse algoritmo, foi usado a função StandardScaler() para padronizar as variáveis de média de peso e altura, pois os valores originais tem escalas bem diferentes, e para o KNN é preciso escalas similares para a comparação.

Tempo de treino	Acurácia	Precisão	Recall	F1-score	Tempo de predição
0.0414 segundos	0.4827	0.4367	0.4827	0.4562	0.0367 segundos

4.3 Algoritmo Random Forest

O último algoritmo escolhido foi o Random Forest. Aqui também foi utilizado a variação da classificação da class/target *obey*.

Tempo de treino	Acurácia	Precisão	Recall	F1-score	Tempo de predição

0.3535 segundos	0.4137	0.3077	0.4137	0.3433	0.0412 segundos

4.4 Comparação entre os resultados

Algoritmo	**Tempo de treino**	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**Tempo de predição**
Árvore sem variação	0.0039 segundos	0.2413	0.2403	0.2413	0.2397	0.0040 segundos
Árvore com variação	0.0108 segundos	0.4137	0.3192	0.4137	0.3563	0.0020 segundos
KNN	0.0414 segundos	0.4827	0.4367	0.4827	0.4562	0.0367 segundos
Random Forest	0.3535 segundos	0.4137	0.3077	0.4137	0.3433	0.0412 segundos

5. Conclusão

Após os testes realizado com os 3 algoritmos e variações, pode-se concluir que o algoritmo que obteve o melhor desempenho foi o KNN. Praticamente todas as suas métricas são superiores aos outros algoritmos. Então, é possível afirmar que, para o objetivo de prever a categoria de obediência de uma raça de cachorro através da altura média e peso médio, ele é a melhor escolha, entre as testadas.

A class que foi melhor prevista foi a “baixa” da class/target, pois há mais exemplos dela no dataset. Ao contrário da classe “alta”, que possui menos exemplos e se torna mais difícil de identificar.

O pior algoritmo foi a Árvore de Decisão antes do pré-processamento. Devido a existência de 5 classes numéricas na class/target, ele teve dificuldade em prevê-las através de apenas 2 atributos (o peso e a altura médios).

O dataset não é grande, o que impactou diretamente no desempenho não muito alto de todos os algoritmos treinados. Ainda sim, foi possível obter resultados satisfatórios com o KNN.

É possível inferir através da análise que as medidas físicas tem alguma relação com obediência, mas não são fortes preditores. Além disso, percebe-se que os algoritmos baseados em distância, como o KNN, funcionam melhor do que os de árvores. As classes com menor frequência, como a “alta” obediência, se tornam mais difíceis de serem previstas.