Trabalho de Aprendizado de Máquina Supervisionado – Descrição das Bases de Dados e Atributos (Entrega Parcial – 26/05/2025)

Integrantes do Grupo

Bruno Franz
Josué Weschenfelder
Engenharia de Software
Inteligência Artificial

Bases de Dados Selecionadas

Para cumprir o requisito de trabalhar com três formatos distintos (tabular, imagem e texto) e avaliar diferentes técnicas de classificação supervisionada, foram escolhidos conjuntos públicos de tamanho médio, rótulos bem definidos e licença aberta para uso acadêmico.

---

1. Base Tabular – Bank Marketing (UCI / Kaggle)

O conjunto Bank Marketing reúne 45 211 registros de chamadas de telemarketing feitas por um banco português entre 2008 e 2010. Cada instância descreve um cliente contactado, e o objetivo é prever se ele subscreve (yes) ou não (no) um depósito a prazo. A base está em CSV e não contém valores ausentes.

Atributos por tipo

- **Numéricos contínuos ou de contagem**: idade (age), saldo médio anual (balance), duração da chamada em segundos (duration), dia do mês da chamada (day), número de contatos na campanha atual (campaign), dias desde o último contato anterior (pdays), total de contatos em campanhas anteriores (previous). Estes campos podem exigir padronização antes do treinamento.
- **Categóricos**: profissão (job), estado civil (marital), nível de educação (education), tipo de contato (contact – telefone fixo ou celular), mês da chamada (month) e resultado da campanha anterior (poutcome). Todos precisam ser codificados (por exemplo, one-hot).
- **Binários**: histórico de inadimplência (default), existência de empréstimo habitacional (housing) e existência de empréstimo pessoal (loan). Já vêm como “yes”/“no” e podem ser convertidos diretamente em 0/1.
- **Rótulo**: campo y, binário, indicando a adesão ao produto financeiro.

---

2. Base de Imagens – Flowers Recognition (Kaggle)

O dataset **Flowers Recognition** contém 4 242 fotografias JPEG (~320 × 240 px), divididas em cinco pastas que correspondem às espécies *daisy*, *dandelion*, *rose*, *sunflower* e *tulip*. O problema de interesse é classificar cada imagem na sua categoria correta.

Atributos por tipo

- **Matriz de pixels RGB**: cada imagem representa uma amostra; é recomendável redimensionar para um tamanho uniforme (p.ex. 224 × 224 px) e normalizar valores de cor antes de alimentar uma CNN.
- **Rótulo**: a espécie da flor, derivada diretamente da pasta que contém a imagem (cinco classes no total).

---

3. Base Textual – Books Reviews em Português (GitHub)

- Books Reviews reúne 2 000 avaliações de livros publicadas por usuários da Amazon Brasil. Metade dos comentários está no arquivo books_pt_neg e foi classificada como negativa (abaixo de 3 estrelas); a outra metade encontra-se em books_pt_pos e corresponde a resenhas positivas (acima de 3 estrelas). O problema é identificar automaticamente se o texto expressa opinião favorável ou desfavorável. Atributos por tipo Texto livre principal: cada linha de texto contém uma resenha completa, que precisa ser tokenizada e vetorizada (TF-IDF ou embeddings) para alimentar os modelos. Rótulo: 0 para books_pt_neg e 1 para books_pt_pos. Como os arquivos não trazem metadados adicionais, o foco é exclusivamente na classificação textual.

Adequação aos Requisitos

- **Diversidade de formato**: inclui dados tabulares, imagens e texto, cobrindo diferentes etapas de pré-processamento e arquitetura de modelos.
- **Separação clara dos tipos de atributo**: facilita planejar normalização, codificação ou tokenização, conforme o caso.
- **Tamanho gerenciável**: cada conjunto cabe em equipamentos de laboratório ou notebooks pessoais, permitindo realizar as três repetições de hiperparâmetros e as cinco arquiteturas de redes neurais exigidas.
- **Fontes públicas confiáveis**: UCI Repository, Kaggle e Github todos com licença aberta para uso acadêmico.

A seguir registramos **os primeiros ensaios de treinamento** em duas das bases, já medindo acurácia, precisão, recall, F-score e tempos de execução. Esses números serão expandidos quando todas as combinações e arquiteturas estiverem prontas.

4.1 Bank Marketing (UCI) – Árvores de Decisão

Config.	Hiperparâmetros	Acurácia	Precisão*	Recall*	F1*	Tempo treino	Tempo predição
1	`criterion="gini"`, `max_depth=None`	**0.886**	0.472	**0.510**	**0.490**	0.33 s	4 ms
2	`criterion="entropy"`, `max_depth=5`	0.901	0.552	0.439	0.489	**0.05 s**	3 ms
3	`criterion="gini"`, `max_depth=10`, `min_samples_split=10`	**0.905**	**0.591**	0.398	0.476	0.08 s	3 ms

* métricas calculadas sobre a classe minoritária “yes”.
Interpretação: a Árvore profunda (Config. 1) obteve o melhor recall e F-score, enquanto a árvore podada a profundidade 10 (Config. 3) maximizou acurácia e precisão. Para um cenário em que falsos-negativos custam mais (perder um cliente potencial), Config. 1 é preferível; caso contrário, Config. 3 oferece ligeiro ganho de acurácia sem penalizar muito o tempo.

https://archive.ics.uci.edu/dataset/222/bank%2Bmarketing?utm_source=chatgpt.com

https://www.kaggle.com/datasets/alxmamaev/flowers-recognition?resource=download

https://github.com/larifeliciana/books-reviews-portuguese