Detailed Report on Data Preprocessing

This report outlines the specific cleaning and transformation steps applied to each of the three datasets. The goal is to convert the raw data into a clean, fully numerical format suitable for the k-Nearest Neighbors algorithm.

---

1. Hayes-Roth Dataset

This dataset involves classifying a person based on personal attributes.

1.1 Original Columns & Cleaning Steps:

- `name`: A unique identifier for the person. **Action:** This column was **dropped** as it provides no predictive information.
- `hobby`: The person's hobby (e.g., sports, reading). Represented as a number from 1-3.
- `age`: The person's age group. Represented as a number from 1-4.
- `educational_level`: The person's level of education. Represented as a number from 1-4.
- `marital_status`: The person's marital status. Represented as a number from 1-4.
- `class`: The target classification for the person. Represented as a number from 1-3.

1.2 Final Processed Columns:

The data was already numerical, so no encoding was needed. The final processed files contain:

- **`X_hayes_processed.csv`**:
  - `hobby`: Integer (1-3)
  - `age`: Integer (1-4)
  - `educational_level`: Integer (1-4)
  - `marital_status`: Integer (1-4)
- **`y_hayes_processed.csv`**:
  - `class`: Integer (1-3)

This data is ready for the **Hamming Distance** metric.

---

2. Car Evaluation Dataset

This dataset classifies the acceptability of a car. All features are categorical and have a natural order.

2.1 Original Columns & Preprocessing (Ordinal Encoding):

The primary task was to convert the text categories into numbers that respect their order.

Human-Readable Name	Original Values	Post-Processing Content (Numerical Value)
buying_price	low, med, high, vhigh	**Ordinal Integer:** 0, 1, 2, 3
maintenance_cost	low, med, high, vhigh	**Ordinal Integer:** 0, 1, 2, 3
doors	2, 3, 4, 5more	**Ordinal Integer:** 0, 1, 2, 3
person_capacity	2, 4, more	**Ordinal Integer:** 0, 1, 2
luggage_boot_size	small, med, big	**Ordinal Integer:** 0, 1, 2
safety_rating	low, med, high	**Ordinal Integer:** 0, 1, 2
class	unacc, acc, good, vgood	**Ordinal Integer:** 0, 1, 2, 3

Export to Sheets

2.2 Final Processed Columns:

The final processed files contain these new numerical representations, allowing for a meaningful distance calculation. This data is ready for the **Manhattan Distance** metric.

---

3. Breast Cancer Dataset

This dataset predicts cancer recurrence based on tumor characteristics. The features are categorical with no natural order.

3.1 Original Columns & Preprocessing (Imputation & One-Hot Encoding):

This was a two-step process: cleaning missing values and then transforming the data.

Human-Readable Name	Original Content	Cleaning / Preprocessing Actions
age	Age bracket (e.g., 30-39)	Converted to multiple binary columns (e.g., age_30-39, age_40-49).
menopause	Menopause status (e.g., premeno)	Converted to multiple binary columns.
tumor-size	Size bracket (e.g., 30-34)	Converted to multiple binary columns.
involved_lymph_nodes	Number of affected nodes	Converted to multiple binary columns.
node-caps	If cancer spread to node capsules	Missing '?' values filled with the **mode**. Then converted to binary column.
malignancy_degree	Pathologist's grade (1-3)	Converted to multiple binary columns.
breast	Left or right breast	Converted to a single binary column.

Human-Readable Name	Original Content	Cleaning / Preprocessing Actions
`tumor_location_quadrant`	Location on breast	Missing `'?'` values filled with the **mode**. Then converted to binary columns.
`received_radiation`	Yes/No	Converted to a single binary column.
`class`	`no-recurrence-events`, `recurrence-events`	Mapped to **`0`** and **`1`**.

Export to Sheets

3.2 Final Processed Columns:

After one-hot encoding, the original 9 feature columns were expanded into 32 new binary (0/1) columns.

- **`X_cancer_processed.csv`**: Contains 32 columns like `age_30-39`, `menopause_premeno`, `tumor-size_30-34`, etc. Each column contains only `0`s and `1`s.
- **`y_cancer_processed.csv`**: Contains the single `class` column with `0`s and `1`s.

This high-dimensional binary data is ready for the **Euclidean Distance** metric.