Feature Name	Type	Description	How It Is Computed	Why It Helps the Task
Declared Value (CAD)	Numeric	The item's declared monetary value.	Parsed from CSV, cleaned, and normalized with z-score scaling.	High-value goods often map to specific HS categories (electronics, jewelry).
Weight (kg)	Numeric	The package's physical weight.	Cleaned, converted to numeric, and normalized.	Heavier goods typically belong to machinery/industrial codes; lighter goods often include apparel/accessories.
Gift Flag	Binary (0/1)	Whether the shipment is marked as a gift.	Extracted as 0 (not gift) or 1 (gift).	Some HS codes treat personal gifts differently; it helps distinguish retail goods from personal items.
Origin Country	Categorical (One-hot)	Country from which the item is shipped.	One-hot encoded into separate boolean columns.	Certain HS categories exhibit a strong correlation with common exporting countries.
Destination Country	Categorical (One-hot)	The country receiving the shipment.	One-hot encoded.	Some destinations require different tariff categories for similar goods.
Category Tag(s)	Categorical (Multi-hot)	High-level product categories are provided in metadata.	Multi-hot encoded vector (each tag = 0/1).	Gives coarse semantic grouping (e.g., "cosmetics", "electronics"). Improves model confidence sharply.
Product Title	Text (raw)	Short product description string.	Passed through tokenization -> vocabulary -> TF-IDF vectorization.	Most predictive field; contains keywords like "lipstick", "charger", "toy", etc.
TF-IDF Word Features	Sparse numeric vector (300–900 dims)	Weighted term frequencies representing the title/description.	TF-IDF transformation using sklearn, with rare-word smoothing.	Captures semantic similarity, which is essential for KNN and decision-tree splits.
Length of Title	Numeric	Number of characters/words in the title.	Computed directly and scaled.	Some product categories correlate with short vs long titles (e.g., "iPhone Case" vs complex chemical names).
Punctuation / Symbol Flags	Binary	Flags for "%", "ml", "cm", "USB", "pack", etc.	Regex-based indicators added to the feature vector.	Certain HS classes (cosmetics, electronics, textiles) use consistent unit abbreviations.
Clean/Noisy Dataset Selector	Binary	Whether an instance comes from the clean or noisy dataset.	Automatically set by the preprocessing pipeline.	Helps the model generalize differently for clean vs noisy samples (optional).