The Necessity for Nutrition

**Anthony Ostia aostia@msn.com**

UC San Diego

La Jolla, CA

**Abstract**

Consumers often prioritize visual appeal and convenience over nutritional quality when selecting foods. This study asks whether popularity correlates with poorer nutrition by analyzing three public datasets—breakfast cereals, McDonald’s menu items, and Starbucks beverages. Each dataset is modelled with logistic regression, support-vector machines, and decision trees under two train–test splits (80 %/20 % and 20 %/80 %). Comparative test-error rates quantify the influence of classifier choice and training-set size on predictive performance. Decision trees deliver the lowest generalization error, and popular items consistently exhibit higher amounts of sugar or fat than their less-advertised counterparts. These findings highlight systematic nutrient imbalances in high-visibility products and underscore the importance of transparent nutrition information in guiding healthier consumer choices.

**1. Introduction**

Nutrition science consistently shows that habitual excesses of added sugar, saturated fat, and calories elevate the risk of obesity and cardiometabolic disease. Yet time-pressed consumers frequently choose foods for speed and marketing appeal rather than nutrient content—a pattern reinforced by the ubiquity of quick-service restaurants and pre-packaged meals. High-profile products from brands such as Kellogg’s, McDonald’s, and Starbucks dominate advertising budgets and shelf space, but the extent to which their nutritional profiles differ from less-promoted alternatives remains under-examined.

This paper investigates whether popularity is a reliable proxy for inferior nutrition. We formulate binary classification tasks on three open datasets: (i) top-selling versus other breakfast cereals, (ii) McDonald’s lunch versus breakfast entrées, and (iii) Starbucks espresso-based versus drip beverages. Logistic regression, support-vector machines, and decision trees are trained under two cross-validation regimes (80 %/20 % and 20 %/80 % train–test splits). By comparing test errors and analyzing feature importances, we quantify nutrient patterns that distinguish high-visibility foods and assess how classifier complexity influences generalization.

**2. Methods**

**2.1 Datasets**

* **Breakfast cereals** – 77 products with nutrient values provided in integers (e.g., calories, sugar).
* **McDonald’s menu** – Entrées from the official nutrition PDF released on Kaggle.
* **Starbucks beverages** – Full drink lineup extracted from Starbucks’ public nutrition brochure.

**2.2 Class Definition**

To examine whether *visibility* aligns with poorer nutrition, each dataset was binarized:

* **Cereals:** two top-selling brands (Class A) vs. all other brands (Class B).
* **McDonald’s:** lunch items (A) vs. breakfast items (B).
* **Starbucks:** espresso-based drinks (A) vs. brewed coffee drinks (B).

**2.3 Feature Selection**

* **Cereals & Starbucks** – calories (kcal) and sugar (g) capture the main consumer health concerns.
* **McDonald’s** – calories (kcal) and total fat (g) reflect the menu’s dominant macronutrient risk.  
  No imputation was required because all datasets contained complete numeric entries.

**2.4 Experimental Partitions**

Two hold-out splits test sensitivity to training-set size:

| **Partition** | **Train: Test** |
| --- | --- |
| **P₁** | 80 %: 20 % |
| **P₂** | 20 %: 80 % |

**2.5 Classifiers**

1. **Logistic Regression** – baseline linear separator mapping log-odds of class membership.
2. **Support-Vector Machine (RBF kernel)** – maximizes the margin via a nonlinear hyper-plane.
3. **Decision Tree (CART)** – recursively splits integer features to minimize Gini impurity.

Each model was trained three times on both partitions; mean training and test errors were recorded.

**3. Experiments**

The experiment proceeds in three stages:

1. **Model Training** – For each dataset × partition × classifier triple, fit the model on the designated training subset.
2. **Performance Evaluation** – Record training error and test error per run, then take the arithmetic mean across three repetitions to mitigate random-seed variance.
3. **Visual Diagnostics** –
   * Heat-maps of decision-tree depth vs. cross-validated error determine optimal max\_depth.
   * 2-D scatter plots overlay the learned decision boundary onto (calories, sugar/fat) space, illustrating class separation.

A screenshot of a computer

AI-generated content may be incorrect.

**5. Conclusion**

Using three public nutrition datasets, we show that model accuracy—and hence the reliability of popularity-vs-nutrition claims—depends strongly on both training-set size and classifier choice. An 80 / 20 split consistently produced lower test error than the more data-sparse 20 / 80 alternative, confirming conventional wisdom that larger training corpora combat over-fitting. Across all datasets, the simple CART decision tree achieved the smallest generalization error, outperforming logistic regression and RBF-SVM despite their higher representational capacity. Qualitatively, the “popular” class (top brands, lunch entrées, espresso drinks) exhibited markedly higher median sugar or fat values, reinforcing concerns that marketing visibility aligns with poorer nutritional profiles. These findings underscore the importance of transparent labelling and suggest that even lightweight machine-learning models can offer actionable insights into consumer health education. Future work should incorporate additional nutrients (e.g., sodium), expand to international menus, and explore ensemble methods to boost robustness.