A Computational Framework for Sustainable Material Selection via Web Scraping, Data Imputation, and Multi-Criteria Decision Analysis

Aayush Shailesh Jhawar (230028), Mechanical Engineering
Sanchit Garg (230908), Mechanical Engineering
Gauri Gupta (230415), Earth Science
Aryan Raj (230219), Earth Science

Indian Institute of Technology, Kanpur

Abstract

The selection of engineering materials has become a critical sustainability challenge, requiring a balance between mechanical performance, cost, and environmental impact. Existing materials databases are either incomplete or not structured for computational decision-making. This work presents a complete, end-to-end pipeline that integrates web scraping, data cleaning, category-weighted imputation, and a hybrid multi-criteria decision analysis (MCDA) framework combining TOPSIS and PCA. A dataset of over 2,400 materials was extracted from MatWeb, standardized, and augmented with CO_2 footprint and cost data. The proposed model demonstrates an interpretable, scalable approach to sustainable material selection for engineering applications.

Keywords: Sustainable Materials, Data Imputation, Web Scraping, TOPSIS, PCA, Multi-Criteria Decision Analysis, CO_2 Footprint

1 Introduction

The accelerating demand for sustainable design has led to the need for materials that balance performance, cost, and environmental impact. However, publicly available data is often fragmented across incompatible sources, with inconsistent units and missing values. Conventional databases such as MatWeb or CES EduPack provide limited machine-readability and rarely integrate life-cycle or cost dimensions.
This work proposes a computational framework to automate the collection, cleaning, and analysis of material data, enabling holistic material selection. The framework combines web scraping, intelligent imputation, and a multi-criteria scoring system integrating engineering, environmental, and economic criteria.

2 Methodology

2.1 Data Acquisition

Material data were extracted using a custom Python scraper built on Playwright to access non-API endpoints from the MatWeb database. The scraper collected over 2,400 material entries covering ceramics, polymers, metals, composites, and glasses. Each record included density, modulus of elasticity, thermal conductivity, and optical/electrical parameters.

2.2 Data Cleaning and Normalization

The scraped data suffered from inconsistent formatting (e.g., mixed units such as g/cc and lb/in³). A regex-based parser and unit normalizer were implemented to convert all properties into SI units. Non-numeric values were filtered, and missing data were imputed through statistical and category-weighted methods.

2.3 Category-Weighted Imputation

To handle missing data, materials were classified based on hierarchical categories (e.g., Ceramic;Oxide;Aluminum Oxide). Each material's missing properties were imputed by computing weighted averages across category levels:

$$x_{\text{imputed}} = 0.6x_{\text{main}} + 0.3x_{\text{sub}} + 0.1x_{\text{extended}}$$

This approach preserves structural differences among material classes, avoiding unrealistic uniformity while filling critical gaps.

2.4 Multi-Criteria Scoring Framework

Each material was evaluated through a hybrid scoring system. Principal Component Analysis (PCA) was used to identify dominant variance directions among material properties. Then, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) ranked materials according to proximity to the ideal solution across normalized, weighted attributes. Cost and CO_2 data were treated as negative criteria (lower is better).

3 Results and Analysis

After preprocessing, 2,456 valid material records were retained. The PCA revealed that 82.4% of the variance was explained by the first three components, dominated by mechanical stiffness, density, and thermal conductivity.

The TOPSIS ranking produced distinct clusters of high-performing materials, with lightweight ceramics and advanced composites scoring highest when normalized across all three dimensions—engineering, environmental, and economic. Metals like titanium and aluminum alloys performed competitively but were penalized for cost and CO_2 impact.

Correlation analysis highlighted significant inverse relationships between density and specific stiffness, and between cost and sustainability scores.

4 Discussion and Evaluation

The integration of environmental and cost attributes within a quantitative MCDA pipeline represents a shift from descriptive to prescriptive material selection. The imputation algorithm effectively balanced data completeness without overfitting to class medians.

However, limitations persist in cases where subcategory representation was sparse, potentially biasing averages. Furthermore, CO_2 and cost data were sourced from secondary repositories (IPCC, Engineering Toolbox), which may not reflect process-specific variations.

5 Conclusion and Future Work

This project demonstrates a modular and reproducible computational framework for sustainable material selection. By integrating data engineering, statistical imputation, and multi-criteria optimization, it bridges the gap between open data and real-world engineering decision-making.

Future work will focus on integrating life-cycle assessment (LCA) databases, automating feature scaling based

on uncertainty, and extending the methodology toward real-time materials informatics systems. The pipeline's design lends itself to publication and potential industrial application in design optimization.

Acknowledgements

We thank Prof. Amar Behera for his guidance throughout this project. The authors also acknowledge the open data resources that made this research possible.

References

1. Ashby, M.F. *Materials Selection in Mechanical Design*. Elsevier, 2017.
2. Hwang, C.L., Yoon, K. (1981). *Multiple Attribute Decision Making: Methods and Applications*. Springer.
3. MatWeb Material Property Data. `https://www.matweb.com`
4. IPCC Emission Factors Database, 2021. `https://www.ipcc.ch/data/`
5. Engineering Toolbox Materials Database. `https://www.engineeringtoolbox.com`