

Unsupervised learning model to determine a classification based on the cost-benefit ratio of school notebooks in Mexico**Modelo de aprendizaje no supervisado para determinar una clasificación basada en la relación costo-beneficio de los cuadernos escolares de México**URUETA-HINOJOSA, Daniel Edahi^{1†*}, GALVÁN-ZAVALA, Karina², LAVÍN-DELGADO, Jorge Enrique¹ and SUÁREZ-ESCOBAR Jorge Armando³¹*Network and Telecommunications Engineering, Universidad Politécnica del Estado de Guerrero, Iguala-Taxco Federal Highway, Kilometer 105, Z.C. 40321*²*Department of management and business management. Universidad de Guanajuato. Lascuráin de Retana #5, Guanajuato, Guanajuato. Z.C.36000*³*International Trade and Customs, Universidad Politécnica del Estado de Guerrero, Iguala-Taxco Federal Highway, Kilometer 105, Z.C. 40321*ID 1st Author: *Daniel Edahi, Urueta-Hinojosa* / **ORC ID:** 0000-0002-8741-6978ID 1st Co-author: *Karina, Galván-Zavala* / **ORC ID:** 0000-0001-5759-8814ID 2nd Co-author: *Jorge Enrique, Lavín-Delgado* / **ORC ID:** 0000-0003-3632-3373ID 3rd Co-author: *Jorge Armando, Suárez-Escobar* / **ORC ID:** 0009-0004-5838-6820**DOI:** 10.35429/EJT.2023.14.7.24.28

Received July 22, 2023; Accepted December 11, 2023

Abstract

One of the most demanded items at the beginning of each school year is the notebook besides, there are different ways of classifying them, the result is usually biased towards their characteristics, not their cost-benefit ratio. Taking studies carried out by PROFECO on the most purchased notebooks in Mexico as a source of reference and comparison, this paper shows as well as compares the results of a proposed unsupervised learning model to classify school notebooks in a cost-benefit ratio.

Comparison, Classification, Notebooks, Unsupervised learning**Resumen**

Uno de los artículos más demandados al inicio de cada ciclo escolar es el cuaderno y aunque existen diferentes formas de clasificarlos, el resultado suele estar sesgado a sus características y no a su relación costo-calidad. Tomando como fuente de referencia y comparación estudios realizados por la PROFECO de los cuadernos más comprados en México, en este trabajo se muestran y comparan los resultados de un modelo propuesto de aprendizaje no supervisado para clasificar los cuadernos basándose en una relación costo-beneficio.

Comparación, clasificación, Cuadernos, aprendizaje no supervisado

Citation: URUETA-HINOJOSA, Daniel Edahi, GALVÁN-ZAVALA, Karina, LAVÍN-DELGADO, Jorge Enrique and SUÁREZ-ESCOBAR Jorge Armando. Unsupervised learning model to determine a classification based on the cost-benefit ratio of school notebooks in Mexico. ECORFAN Journal-Taiwan. 2023. 7-14: 24-28

* Author's Correspondence (e-mail: daniel.urueta@upeg.edu.mx)

† Researcher contributing as first author.

Introduction

Every school year, Mexican parents must spend an approximate amount of 53% of their monthly income (WorldRemit, 2022); according to studies, the most demanded item is the notebook; however, their average prices have increased by 123% compared to the previous year (Valladolid, 2023). Therefore, it is necessary to have tools that allow the best possible decision to be made with an increasingly tight budget.

One of these tools is the analysis developed by the Federal Consumer Attorney's Office (PROFECO, 2023), which is in charge of make quality studies about multiple products along with brands available in Mexican market in addition to report them publicly; nonetheless, it issues recommendations so that consumers can establish comparisons as well as make the best purchase decisions. Nonetheless, frequently these studies, as well as all those carried out by PROFECO, are based only on a descriptive classification more than on a formal statistical analysis to determine the best products considering the cost-benefit ratio.

This paper describes the results of a set of techniques based on unsupervised learning to classify school notebooks in Mexico according to their cost-benefit ratio in addition to taking into account the studies carried out by PROFECO.

Objectives

General

- Determining through artificial intelligence techniques the best school notebooks using their cost-benefit ratio.

Specifics

- Complementing the studies carried out by PROFECO.
- Helping the consumer to take better decisions when they have a tight budget.

State of art

Unsupervised learning

Within artificial intelligence, there are algorithms which do not need to be trained in order to classify a dataset; conversely, their classification is done through the inherent characteristics of each element, this type of technique is called unsupervised learning classification or clustering.

The techniques used are particularly useful when there is no prior labeling, that is, when it is unknown how a dataset can be organized, notwithstanding, it is necessary classifying it. Furthermore, these techniques are often used for comparison purposes as well as to get feedback between the results (classification) produced by a model against those determined by a specialist (Urueta *et al.*, 2021).

K-means

Perhaps, this is the best-known clustering technique, its operation consists of splitting the data into a number k of groups. Each data, also called observation, is assigned to one group or another depending on how close the observation is to the average value of each group.

Elbow method

A problem with the k -means method is that it is necessary to specify the number k of groups into which the database will be separated beforehand; to accomplish this task, the elbow method was used.

The elbow method is a measure to calculate the sum of squared error (SSE) between the mean distance and all other data points in the same group. It can formally be defined as:

$$SSE = \sum_{i=1}^n (x_i - \bar{x})^2 \quad 1$$

It is needed to calculate the SSE values and then plotted them from 2 to n ; thus, if we show the graphic, the point where the SSE or inertia values start decreasing in a linear manner will be the recommended k value to classify the database.

Principal Component Analysis: PCA

This technique allows to reduce the dimensions of a dataset, trying to preserve the greatest variation. It works by searching for possible correlations by means of lines, planes, and hyperplanes in the original space.

Pearson correlation coefficient

It is a measure which determines how two variables are linearly related. This coefficient is represented by the letter *r*. The result will be from -1 to +1. It can be interpreted as in Figure 1.

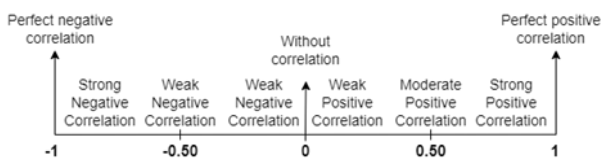


Figure 1 Interpreting a correlation analysis

PROFECO

The Federal Consumer Attorney's Office, abbreviated as PROFECO in Spanish, seeks to empower the consumer through the effective protection of the exercise of their rights and citizen trust, promoting a reasoned, informed, sustainable, safe as well as healthy consumption, in order to correct market injustices, strengthen the internal market along with the well-being of the population (e.Economía, 2022). In this sense, PROFECO has the National Consumer Protection Laboratory (LNPC, by its acronym in Spanish), dedicated to preparing as well as publishing quality studies of widespread consumption products on their impact on nutrition, safety along with economy (AMEXCID, 2018).

Materials and methods

Materials

- For this study, the PROFECO databases were used, this can be consulted from (PROFECO, 2022).
- Python programming language for database analysis.

Methods

- Normalize the database
- Apply artificial intelligence techniques, such as: PCA in addition to unsupervised classification methods to the normalized database.
- Compare the results obtained by the model with those reported by PROFECO.
- Report the results in a clear way.

Results

Pearson correlation analysis applied on the original database (see Figure 2) demonstrates the following:

- There is a moderate positive correlation between the cost with the grams of sheets indicated and the real one.
- There is a moderate positive correlation between the classification proposed by PROFECO and the resistance of the sheets and cover.
- There is no correlation between the cost and the classification proposed by PROFECO.

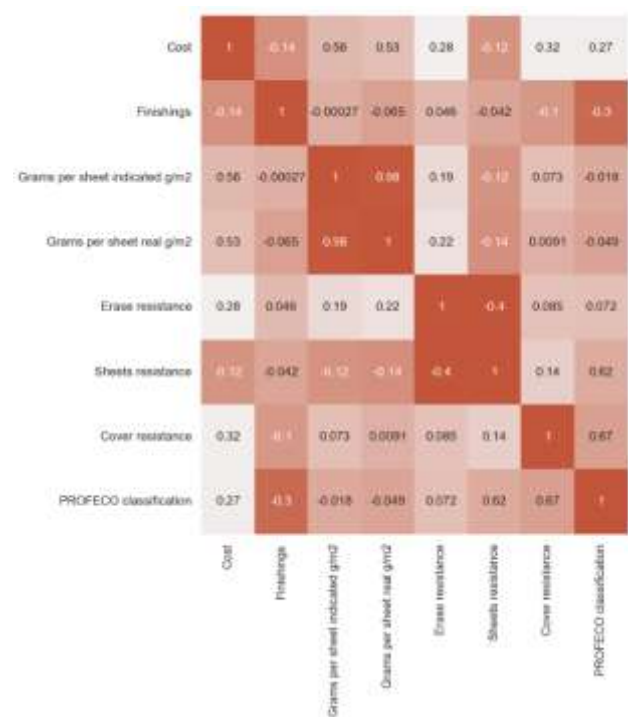


Figure 2 Correlation analysis with original classification

Subsequently, the elbow method was applied on the database to determine the number of classes recommended to classify the dataset.

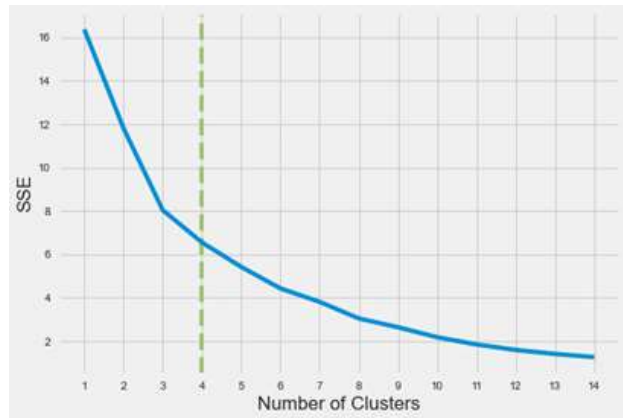


Figure 3 Results of the elbow method

Once the number of clusters was determined (see Figure 3), the database was classified using *k*-means with a value *k* = 4. The classification obtained was added to the database, afterward a correlation analysis was made again. The results of this analysis are presented in Figure 4.

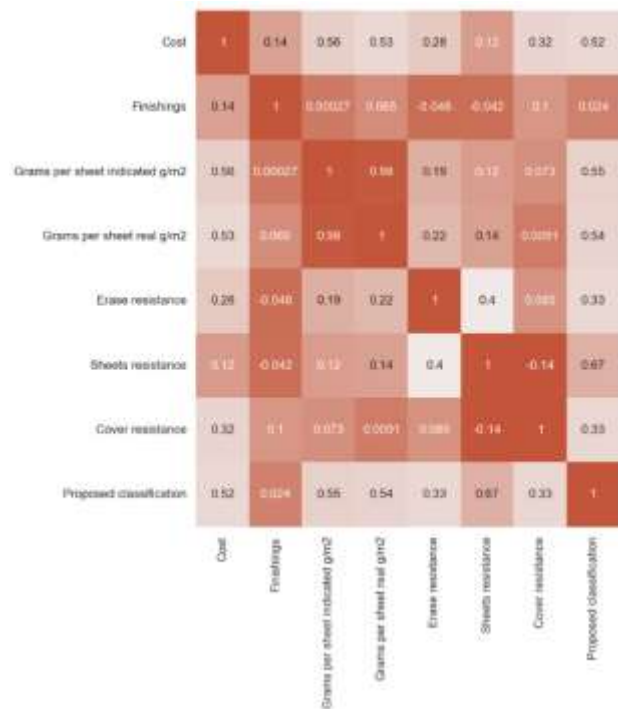


Figure 4 Correlation analysis with the proposed model classification

From the above results, the following can be inferred:

- There is a moderate positive correlation between the classification generated and the cost.

- There is a moderate positive correlation between the classification generated and the grams of indicated sheets.
- There is a moderate positive correlation between the classification generated and the real grams of sheets.
- There is a moderate positive correlation between the classification generated and the resistance of the sheets.

As a result, Table 1 shows a comparison between the qualification assigned by PROFECO against the carried out by the proposed model which considers the cost-benefit ratio.

Model	Qualification Assigned by PROFECO	Classification of the Proposed Model
Scribe 7510	Very good	Excellent
Estrella Único	Very good	Excellent
Manufacturas 8-A	Very good	Excellent
Estrella/ Star Kid	Good	Excellent
Sazz	Sufficient	Sufficient
U-Pak	Sufficient	Sufficient
Norma Cosido	Excellent	Excellent
Estrella Cosido	Excellent	Excellent
Office Max Espiral	Very good	Excellent
Scribe Excellence	Very good	Good
Scribe In Black	Very good	Very good
Manufacturas 8-A	Very good	Very good
First Class 5038	Very good	Very good
First Class 5040	Very good	Very good
First Class 4517	Very good	Very good
Norma Jean Book	Very good	Very good
Scribe Clasico Cosido	Very good	Excellent
Scribe Kids Cosido	Very good	Excellent
First Class 4507	Good	Very good
Estrella Drive Linea A	Good	Good
Scribe In Colors 1070	Good	Very good
First Class Cuaderno Cosido 9408	Good	Sufficient
First Class Profesional 5039	Good	Sufficient
Scribe Poly Cover 2050	Good	Very good
Norma	Good	Sufficient
Estrella 0520	Good	Sufficient
First Class Cosido Profesional 9214	Good	Sufficient
Norma Click Urbana	Good	Sufficient
U-Pak 1157525405	Good	Sufficient
Norma Kiut	Good	Sufficient
Norma Girly Notes	Sufficient	Sufficient
Norma Uno	Sufficient	Sufficient
Nine To Five	Sufficient	Sufficient

Table 1 Comparisons between the PROFECO classification against the proposed model results

Finally, for a better visualization of the classified notebooks, a scatterplot was made reducing the dimensions of the database to 3 by means of PCA. In this scatterplot, the points with the same color mean that they belong to the same class (see Figure 5).

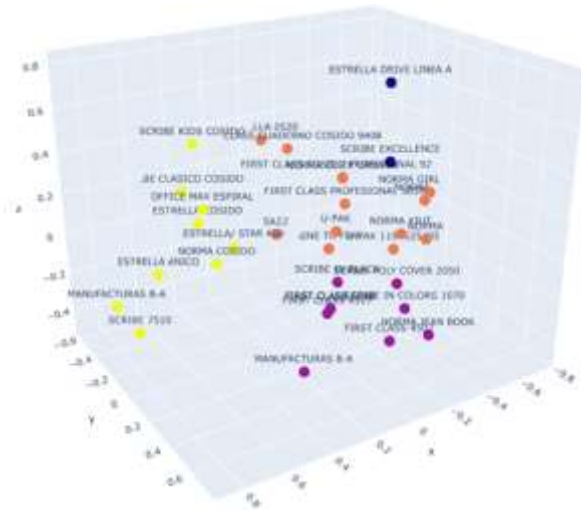


Figure 5 Scatterplot generated with PCA

Acknowledgment

To PROFECO for its studies published about school notebooks in Mexico.

Conclusions

The aim of this paper was to propose an unsupervised learning model classifying notebooks in a cost-benefit ratio, based on the studies accomplished by PROFECO on the most purchased notebooks in Mexico.

The proposed model considers both the cost and the characteristics of the notebooks, such as the number of sheets, the quality of the paper, the type of binding, and the design of the cover. The classification performed by PROFECO, on the other hand, focuses more on the resistance of the sheets and covers, which may not reflect the actual needs or preferences of the consumers.

One of the challenges that we faced was to obtain reliable and updated data on the prices along with features of the notebooks in the Mexican market. We used web scraping techniques to collect data from online sources, nevertheless we acknowledge that this may not capture the full range of notebooks available or their actual prices in different regions. Future work could involve conducting surveys or interviews for consumers together with sellers to obtain more accurate and comprehensive data.

This paper provides a novel and useful tool for consumers to establish comparisons and, consequently, buy notebooks that offer the best value for their money.

It also offers insights for sellers to improve their marketing strategies and product quality based on the cost-benefit analysis. Moreover, it contributes to the field of unsupervised learning by applying a clustering algorithm to a real-world problem that has not been explored before. We hope that this paper will inspire further research as well as innovation in this area.

References

- AMEXCID. (2018). *Catálogo de Capacidades Mexicanas de Cooperación Internacional para el Desarrollo*. <https://de.sre.gob.mx/capacidades/objetivos-de-desarrollo-sostenible/producci%C3%B3n-y-consumo-responsables/profeco/estudios-de-calidad>
- e.Economía. (2022). Procuraduría Federal del Consumidor—Desconcentrados [Secretaría de Economía]. *Secretaría de Economía*. <https://e.economia.gob.mx/procuraduria-federal-del-consumidor/>
- PROFECO. (2022). *Útiles escolares*. https://www.gob.mx/cms/uploads/attachment/file/752527/ESTUDIO_DE_CALIDAD_UTILES_ESCOLARES.pdf
- PROFECO. (2023, August 14). *Procuraduría Federal del Consumidor | Gobierno | gob.mx*. <https://www.gob.mx/profeco>
- Urueta, D. E., Lara, P., Gutiérrez, M. Á., de-los-Cobos, S. G., Rincón, E. A., & Mora, R. A. (2021). A Comparative Ranking Model among Mexican Universities Using Pattern Recognition. *Mathematics*, 9(14), 1615. <https://doi.org/10.3390/math9141615>
- Valladolid, M. (2023, January 2). *Cuadernos, el artículo escolar más demandado*. Forbes México. <https://www.forbes.com.mx/cuadernos-el-articulo-escolar-mas-demandado-su-precio-subio-123-en-este-regreso-a-clases>
- WorldRemit. (2022, September 2). *¡Lista de útiles escolares para el regreso a clases!* Blog | WorldRemit. <https://www.worldremit.com/en/blog/spanish/lista-de-utiles-escolares-prepara-a-tus-hijos-para-el-regreso-a-clases-2022/>