Machine learning for libraries with Python libraries: practical case in the Library of Congress of Chile

Marcelo Lorca González [1]
mlorca@bcn.cl ORCID: 0009--0005-6052-9197

Received: 15 June 2024
Accepted: 31 August 2024

This document focuses on the area of machine learning from data, applied to internal processes of a library. This is practical work associated with the development of an application in Python that uses libraries developed for automated learning work. An unsupervised data analysis methodology was applied, called the K-means method (K-medias in Spanish), which allows the data to be segmented or classified into groups to extract common characteristics. Data associated with the library collections were used. The developed code is shared, and visualisations of the data are shown.

Keywords: artificial intelligence, machine learning, clustering, classification, segmentation, k-means, unsupervised data

1 Introduction

The development of artificial intelligence (AI) and machine learning has promoted the emergence of new methodologies in data analysis. On the other hand, the need for information units and libraries to deepen observation of the large amount of data they possess provides a development opportunity for data analysis. Being able to support decision-making with quality data is a very important area, where data analytics acquires great relevance (Minderest 2023).

In the Library of the National Congress of Chile, promoting the use of data in the development of collections is an area that has not been explored much, "although very necessary with the aim of incorporating relevant resources for the parliamentary community." Together with the cataloguing specialists, we decided to work on a data analysis methodology, such as the K-means method, to be applied to one of the most relevant collections of the libraries, such as the monographs of the Valparaíso headquarters. This analysis was applied to bibliographic resources that corresponded to around 30 000 titles. The variety of work to be developed with this data is very wide, including predictive analysis, classification, and cluster analysis. This time, we focused on a segmentation analysis known as K-means, which allowed us to segment our collection according to some characteristic; in this case, the thematic areas that comprise it. For the analysis, the first two digits of the universal decimal classification (UDC) were applied to the titles that were used, taken from left to right. This allowed us to determine the thematic distribution of the collection, establishing more and less relevant areas in it (Boman 2019). The important thing about this analysis was its ability to apply to the entire collection, as well as to specific sectors of it, which allowed us to have both a global and specific vision of our bibliographic data. This practical work was accompanied by images that reinforce its visualisation.

The document is organised into several parts. The first explains concepts about AI and machine learning, programming, Python and machine learning libraries. Then the platform on which the application is programmed is explained. Next, the problem raised, and the proposed solution are explained, applying K-means data analysis. Finally, the developed solution is explained, providing part of the developed code and the visualisations obtained. This work is part of the constant concern on the part of our library to promote the use of new technologies in the development of innovative solutions that provide higher levels of quality to the internal processes of library work. We hope to continue applying other types of studies to data relevant to decision-making, associated with library processes or supporting parliamentary work.

2 Conceptual framework

Explanations are provided on some concepts involved in developing the solution. Currently, there is an explosion in applications associated with AI associated with the type of generative AI, especially due to the emergence of ChatGPT. However, there are also other areas in which this type of applications is being developed. In the area of libraries, there has been very rapid evolution in the development of new technologies in the face of the challenges implied by paradigm changes in the area. On the other hand, the amount of data that libraries work with constitutes a fertile field to experiment and generate new applications using AI. The most relevant concepts of the area defined below:

1. Marcelo Lorca González is Librarian at the Library of the National Congress of Chile (Biblioteca del Congreso Nacional de Chile)

Artificial intelligence concept

Talking about AI has become very common today. The impact of some of these new tools on everyone's lives has led to much interest in this field. There are many concepts and definitions regarding this topic.

For the purposes of this work, we use a definition given by Oracle(n.d.) on their site:

i) Applications that perform complex tasks that previously required human intervention, such as communicating online with customers or playing chess.
ii) Relationship with machine learning.
iii) Machine learning focuses on creating systems that learn or improve their performance based on the data they consume.
iv) It is important to keep in mind that, although all machine learning is AI, not all AI is machine learning.

Machine learning concept

Some characteristic elements of machine learning are the following:

- Ability of machines to generate automatic learning from data.
- Machines identify complex patterns from the data.
- Machines generate predictions.
- Machine use statistical elements.
- According to Ravshan and Umidjon (2024:1) “Machine learning is a technique that trains a system to solve a problem, instead of the solution itself. To solve a problem using Machine Learning, we need to collect outcome data and train a statistical model. Statistical models are mathematically formalized ways of predicting the behavior of a phenomenon.”

1) Python programming language

The basis for developing machine learning applications is programming languages, in particular the following two that stand out:

- Python programming language
- R programming language

Of these, we chose Python to carry out our practices. Python is a powerful and easy-to-learn programming language. It has efficient high-level data structures and a simple but effective object-oriented programming system. The most important characteristics of this language are:

- Interpreted language. This means that it is compiled when the command lines are executed. That is, there are no executable installation files.
- Integration with various operating systems. It can be run from various platforms.
- Large number of libraries in the standard version.
- Large development communities. That is, better processes are generated with the contribution of community members.
- Powerful libraries for AI. These libraries interact with each other; that is, they integrate.

2) Python libraries

- Numpy: This allows Python developers to quickly perform a wide variety of numerical calculations.
- Pandas: A high-level data manipulation tool developed by Wes McKinney. It is built on top of Numpy and enables data analysis that has the data structures we needed to clean the raw data and make it suitable for analysis (e.g., tables). (Zapata 2024)
- Matplotlib: Matplotlib is a comprehensive library for creating static, animated and interactive visualizations in Python. https://matplotlib.org/ (Matplotlib)
- Scikit-learn: Tools for predictive data analysis. This is accessible to everyone and reusable in various contexts. Built on NumPy, SciPy and Matplotlib Open Source, commercially usable: BSD licence. (Scikit-learn n.d.)

3) Machine learning: Types of analysis

The definition is given by Sánchez (2021) to understand what supervised and unsupervised learning is:

> *In the field of Machine Learning there are two important learning paradigms, supervised and unsupervised. The main difference is that supervised learning requires prior knowledge or a label that indicates the output value that our sample should have, this label is known as 'ground truth' (GT). Therefore, the goal of supervised learning is to find a function*

that, given a sample of training data and its respective GT labels, best approximates the relationship between the input and output variables observed in the data. Unsupervised learning, on the other hand, does not depend on output labels, the goal is to find the "intrinsic" structure present in the data set. Supervised learning is often used for classification and regression problems. Some of the most common algorithms are logistic regression, support vector machines, random forests and artificial neural networks. Unsupervised learning is often used for clustering, segmentation and dimensionality reduction, some of the most common algorithms are k-means, principal component analysis (PCA) and non-negative matrix factorization (NMF). (Sánchez 2021).

4) K-means

K-means is an unsupervised classification (clustering) algorithm that groups objects into k groups based on their characteristics. Clustering is performed by minimising the sum of distances between each object and the centroid of its group or cluster. The quadratic distance is usually used. The algorithm consists of three steps:

- Initialisation: Once the number of groups (k) has been chosen, k centroids are established in the data space, for example, by choosing them randomly.
- Assigning objects to centroids: Each data object is assigned to its closest centroid.
- Centroid update: The position of the centroid of each group is updated, taking as the new centroid the position of the average of the objects belonging to the said group. Steps 2 and 3 are repeated until the centroids do not move or move below a threshold distance in each step. The k-means algorithm solves an optimisation problem, with the function to be optimised (minimised) being the sum of the squared distances of each object to the centroid of its cluster (Sinaga & Yang 2020).

5) Problem to solve

Define the most relevant thematic areas of the Valparaíso Monographs collection of the Library of Congress. Based on the analysis of the data, it was decided to obtain the most relevant thematic areas associated with said collection. We associated the topic with the first two digits of the classification number. The classification used in the Library of Congress of Chile is the UDC. By obtaining the first two digits of this classification, we obtained the most important thematic areas. If we add to this the totals of documents associated with them and apply the K-means method, we can segment the thematic areas according to the degree of relevance in the collection. In addition, this study can be replicated to thematic sub-areas of said collection, which allowed us to have a much more complete knowledge of it. This method could also be applied to other collections and documents of the library (Sinaga & Yang 2020).

6) Solution proposal

- The classification number is broken down into the first two digits. The data are cleaned, null values are removed and samples are extracted and analysed.
- Graphic visualisation is generated that distributes the samples in a plane, with the data grouped around the centroids.
- This is complemented by graphics that help understand cases in the visualisation.

7) Algorithm and solution code

- Platform: Colab (Google).
- There are a variety of software that provide us with a work environment or framework with which we can develop our data analysis applications. Of the most famous are Anaconda and Google Colab. In our project, we used Google Colab, a platform that allowed us to create applications with big data and machine learning technology. Python integrates quite well with Google Colab, from where we can obtain varied and powerful features, such as adding new libraries or new tools to exploit data. Google Colab also allowed us to integrate with Google Drive to save our solutions.
- Start algorithm
- We then connected to Google Drive, where we created an application in Colab, which served as a basis for developing the algorithm.
- *Programación de la Solución en Python.*
- Instantiate Libraries (activate the libraries to be used in the application).

```
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import matplotlib.pyplot as plt
```

- Load files into the Colab environment (the Excel file was loaded before being used in the next step).

```
from google.colab import files
uploaded = files.upload()
for fn in uploaded.keys(): print('User uploaded file "{name}" with length {length} bytes'.format (name=fn, length=len(uploaded[fn])))
```

• Data loading.

```
df = pd.read_excel('listadoParaSimposium-SV.xlsx')
df.head(100)
```

- Data preparation.
    1. Data cleaning
        i. to. (filter data)
    2. Processing of null data
    3. Blank data processing

```
df_tematicas_clasif_is_signo_mas = df_tematicas[df_tematicas["Número Clasificación"].notnull()]
df_tematicas_not_null["clasif3"]= df_tematicas_not_null['Número Clasificación'].str.slice(0, 2)
```

    4. Adjust data types

```
df['Materia (tag 650)'] = df['Materia (tag 650)'].astype('string')
df['Número Clasificación'] = df['Número Clasificación'].astype('string')
df['Resumen'] = df['Resumen'].astype('string')
```

- Information about the dataframe (fields and data types)

```
df.info()
```

- Create DF with thematic fields

```
df_tematicas = df[['Número Clasificación', 'Materia (tag 650)']]
df_tematicas.head()
```

1. Filtrar nulos

```
df_tematicas_not_null = df_tematicas[df_tematicas["Número Clasificación"]].notnull()]
```

- Create column classif3, with the first three digits of the classification.

```
df_tematicas_not_null["clasif3"]= df_tematicas_not_null['Número Clasificación'].str.slice(0, 2)
df_tematicas_not_null
```

- I create a method to generate the group by.

```
def my_agg(x):
names = {'Total': x['clasif3'].count()}
 return pd.Series(names)
```

- I deleted index and created column with group by

```
df_tematicas.reset_index(drop=True, inplace=True)
df_tematicas = df_tematicas_not_null.groupby(['clasif3']).apply(my_agg)
df_tematicas
```

- Create df with sort index that I used to display k-means.

```
df_tot_clasif_or['clasif_ind'] = np.arange(len(df_tot_clasif_or))
df_tot_clasif_or.columns
df_tematicas['clasif_ind'] = np.arange(len(df_tot_clasif_or))
df_tematicas.columns
df_clasif_or = pd.DataFrame(df_tematicas)
```

• K-means: We applied the k-means method to be able to visualise the segmentation by subject. That is, the relevance of the thematic areas in the collection.

```
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
import numpy as np
```

```
escalar=MinMaxScaler().fit(df_clasif_or.values)
#escalar
clientes = pd.DataFrame(escalar.transform(df_clasif_or.values),
columns=["Total"])
kmeans = KMeans(n_clusters=3).fit(clientes.values)
clientes["cluster"] = kmeans.labels_
```

- Data visualisation. With this code part we can generate the pie chart visualisations that will complement the results.

```
plt.figure(figsize=(6, 5), dpi=100)
colores = ["red", "blue", "orange", "black", "purple", "pink", "brown"]
for cluster in range(kmeans.n_clusters):
plt.scatter(clientes[clientes["cluster"] == cluster]["clasif_ind"],
clientes[clientes["cluster"] == cluster]["Total"],
marker="o", s=100, color=colores[cluster], alpha=0.2)
plt.scatter(kmeans.cluster_centers_[cluster][0],
kmeans.cluster_centers_[cluster][1],
marker="P", s=280, color=colores[cluster])
plt.title("Distribución Clasif CDU en Colección Valpo", fontsize=20)
plt.xlabel("Clasif CDU", fontsize=15)
plt.ylabel("Total", fontsize=15)
plt.text(1.15, 0.2, "K = %i" % kmeans.n_clusters, fontsize=25)
plt.text(1.15, 0, "Inercia = %0.2f" % kmeans.inertia_, fontsize=25)
plt.xlim(-0.1, 1.1)
plt.ylim(-0.1, 1.1)
plt.show()
```

If everything is executed correctly, we should see something similar to this.

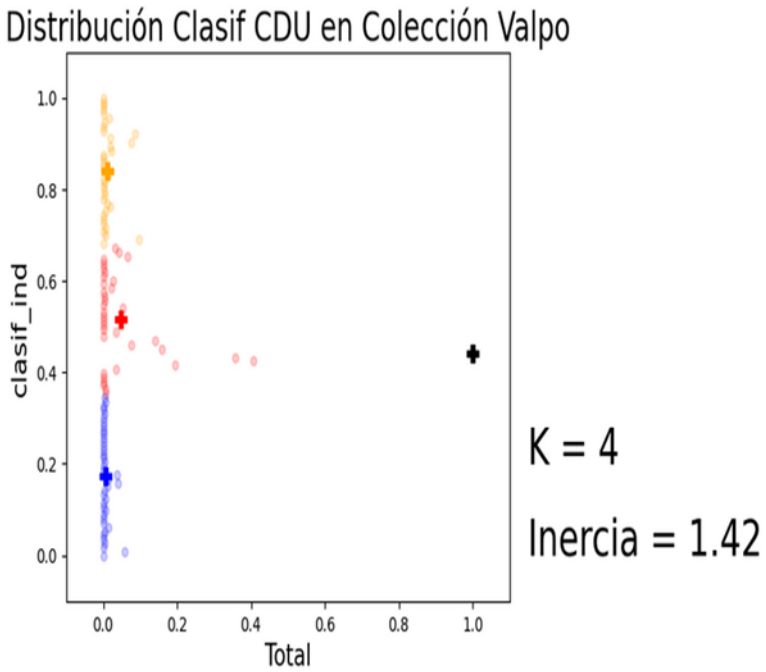

Figure 1:

Figure 1: Resultado aplicación Método K-means. Las escalas usadas reemp´lazan los valores verdaderos, para efectos de visualización. La columna clasif_ind con valores de 0 a 1 representa las agrupaciones generales de la Clasificación Decimal Universal(UDC). La columna del Total representa los valores para cada agrupación, también valores en escala (Result of applying the K-means method. The scales used replace the true values, for visualization purposes. The clasif_ind column with values from 0 to 1 represents the general groupings of the Universal Decimal Classification (UDC). The Total column represents the values for each grouping, also scaled values).

- Elbow method (13)
- This method helped us define the value of K for our K-means. This value allowed us to define how many groups our analysis would include. In effect, the graph had an area where the curve changed direction. The value we took was expected to be in this range. The final value selection was at the discretion of the person doing the analysis, according to the context of the data and the analysis.
- In our practice, through this method, we determined what would be the appropriate number of groups for our analysis. For this practice and seeing that the curve changed direction between 3 and 5, we chose the value 4 as K.

```
inercias = []
for k in range(2, 10):
kmeans = KMeans(n_clusters=k).fit(df_clasif_or.values)
inercias.append(kmeans.inertia_)
plt.figure(figsize=(6, 5), dpi=100)
plt.scatter(range(2, 10), inercias, marker="o", s=180, color="purple")
plt.xlabel("Número de Clusters", fontsize=25)
plt.ylabel("Inercia", fontsize=25)
plt.show()
```

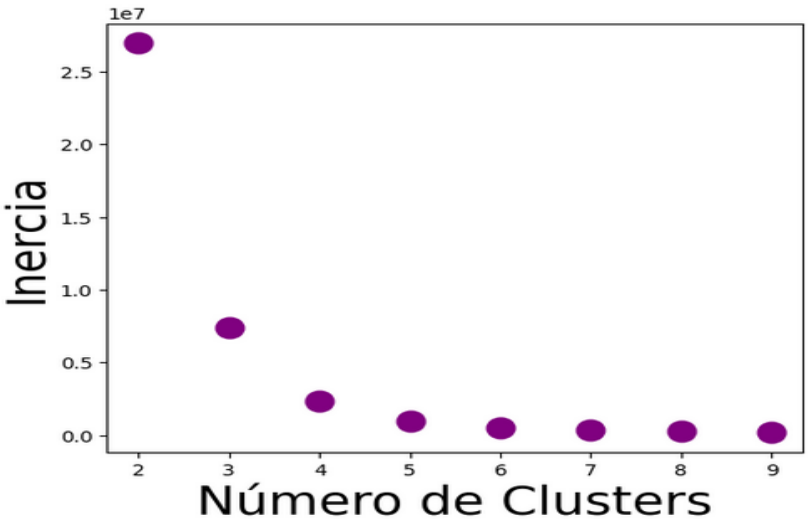

Figure 2:

Figure 2: Visualización Método del Codo. La linea imaginaria que podemos imaginar entre los puntos, nos muestra que el sentido de la misma cambia mayormente entre 3 y 5 (Elbow Method Visualization. The imaginary line that we can imagine between the points shows us that the direction of the line changes mostly between 3 and 5).

3 Conclusion

We carried out a practice applying a type of unstructured data analysis related to the segmentation of titles from our collections, based on our thematic areas. We could replicate this using other types of analysis associated with projections and that allowed us to detect future trends in different aspects of our internal processes. The potential that these tools provide was relevant if we train the results we obtained properly and interpret them judiciously. Indeed, it is important to keep in mind that these data are not absolute truths; their proper interpretation requires placing ourselves in their context and the objective we pursue with our analyses.

The scalability of these data analysis techniques allowed us to apply these works, to achieve knowledge at both a generic level and a specific level of our processes. In this particular case, the analysis applied to see the relevance of the thematic areas of the collection could be applied to only one of these specific areas. This would allow us to achieve a

comprehensive knowledge of our collections at a thematic level. Other types of analysis would allow us to mix this knowledge with other types of methodology data analysis.

Main thematic areas of the UDC

These are the main thematic areas of the UDC, associated with the areas of social sciences. The majority contents of our Valparaíso collection belong to these areas. Area 340 was the main area, as it contained just over 8 000 titles. It was the area that in K-means clustering appeared with only one element in the subset.

i) Thematic princes UDC (300 -399)

1. 310 Statistics
2. 320 Political science
3. 330 Economy
4. **340 Law**
5. 350 Public Administration and Military Science
6. 360 Social problems and services, associations
7. 370 Education

References

Boman, C. 2019. An exploration of machine learning in libraries. In: *Artificial intelligence and machine learning in libraries.* Edited by Jason Griffey. Library Technology Reports alatechsource.org.

Elbow Method for optimal value of k in KMeans. 2023. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/.

IBM. *¿Qué es el aprendizaje no supervisado?* https://www.ibm.com/es-es/topics/unsupervised-learning.

La Python Software Foundation es una organización sin fines de lucro. 2024a. *El tutorial de Python*. Python Software Foundation. https://docs.python.org/es/3/tutorial/.

La Python Software Foundation es una organización sin fines de lucro.2024b. *La biblioteca estándar de Python*. Python Software Foundation. https://docs.python.org/es/3/library/index.html#library-index.

Matplotlib. n.d. *Visualization with Python*. https://matplotlib.org/.

Minderest. 2023. *Qué es la calidad de los datos y cómo te afecta*. https://www.minderest.com/es/blog/calidad-de-datos.

Oracle. n.d. *What is AI? Know artificial intelligence.* https://www.oracle.com/cl/artificial-intelligence/what-is-ai/.

Ravshan, Z. and Umidjon, A. 2024. Building and predicting a neural network in python. *E3S Web of Conferences*, 508(04005), 1-7.

Sánchez, P.A.S. 2021. Métodos de Aprendizaje Supervisado y no Supervisado para la Estimación de Microestructura Cerebral en Datos de DWMR. In: *Centro de Investigación en Matemáticas (CIMAT)* [Preprint].

Scikit-learn. n.d. *Machine learning in Python*. https://scikit-learn.org/stable/.

Sinaga, K.P. and Yang, M.-S. 2020. Unsupervised K-means clustering algorithm. *IEEE Access*, 8: 80716–80727. Available at: https://doi.org/10.1109/ACCESS.2020.2988796.

Zapata, José R. 2024. *Pandas – Manipulación de Datos con Python*. https://joserzapata.github.io/courses/python-ciencia-datos/pandas.