Feature Engineering with Computational Methods for Machine Learning in Remote Sensing Applications

Camilo de la Hoz Lozano
Computational Physics Project

November 2022

Abstract

In this paper, is presented an approach on Feature Engineering for applications on Machine learning in Remote Sensing. Singular Value Decomposition and Principal Component Analysis were applied over a group of bands in order to reduce the amount of data and dimensionality of it. The results provided demonstrate the relevance of Feature Engineering in the field of computer sciences - ML, as it improves drastically the results of a model without changing the parameters of the model or the characteristics of the data-set.

Key words: Singular Value Decomposition, Principal Component Analysis, Machine Learning, Random Forest Classifier, Feature Engineering, Remote Sensing

1 Introduction

Machine Learning (ML) is about the training data-set and the problem to being solved (referring to the question: What is the target and how can I find it?). Having said that, when implementing ML models it is paramount to take some time at analysing the problem and simplifying it by asking the simplest questions and using the best data-set possible (quantity and quality). This process is called Featuring Engineering (FE), and it consist in using previous knowledge in the field of study in order to help the ML model to perform superior and faster (De La Hoz, 2020)

Applications in Remote Sensing (RM) have been increasing as more sensors with superior characteristics are available and the capacities in computing and storage become bigger and accessible. Particularly, among researchers and industry ML is replacing conventional processes for RM land cover classification problems. Here is where FE plays as key factor as it introduces a way of simplifying problems and save on processing time. Specially when data-sets are becoming bigger and include more complex measurements (Anderson et al, 2013).

This approach deals with the question: is Principal Component Analysis (PCA) a good proceed towards FE for improving Supervised Machine Learning problems?. The main objective is determining if through Singular Value Decomposition (SVD) and PCA is it possible to select fewer bands (or a better combination) from a Sentinel-2 raster image and generate better results for Supervised Machine Learning rather than the classic approach in RGB. The hypothesis behind this objective lies in the experiences learned from use of SVD to represent characteristics from a data-set (Barrett & Beroza, 2014) and the well known data reduction method PCA (Jensen, 2005).

2 Area of Study and Data-Set

The study area is located over the University of Alberta (Edmonton, Canada), and its surroundings (see figure 1). The area of study is 14233000 m^2 and 7 bands (out of 16) where selected (selection based on spectral combination and indexes more often used in RM). The image was generated in Google Earth Engine (GEE) and its source code can be found in the following reference: (De La Hoz, 2022).

Figure 1: RGB view of the area of study. the image was taken on the date 2022/08/09 from the Copernicus, Sentinel−2 constellation (*id:* $COPERNICUS/S2/20220809T183931_20220809T184842_T11UQV$).

Selected bands for the study are shown in table 1. This selection was based on previous experience in RM processing for Land-Cover classification.

Band name	Wavelength (nm)	Resolution (m)	Description
B2	496.6	10	Blue
B3	560	10	Green
B4	664.5	10	Red
B8	835.1	10	NIR
B8A	864.8	20	Red Edge 4
B11	1613.7	20	SWIR 1
B12	2202.4	20	SWIR 2

Table 1: Basic information of the bands used for the present study. This information correlates the used information to the original Sentinel−2 image.

The training data-set is composed of a shape file with 3 categories and 15 examples for each. The first one is water, the second is plant and the last one is land (this category takes together soil and men-made land cover). Both the multi-spectral raster image and the vector file can be found at the following link: Public Repository Data-set. Pre-processing of the image (creation of images RGB and PCA selected bands) and vector file (training data-set) were made using **QGIS**. The same link provided leads to the *jupyter notebook* used to perform all the calculations.

3 Methods and Theoretical Background

This study uses three main algorithms. Each one would be discussed and presented as mathematical and theoretical introduction for the method proposed.

3.a Singular Value Decomposition *SVD*

In order to represent each band of the Sentinel-2 image into a smaller set of values rather than a matrix, *SVD* was performed as it decomposes a matrix into a set of vectors (*Singular Vectors of the Matrix*) that contain the most representative characteristics (De La Hoz, 2018). Each band is represented as a matrix A with m rows and n columns. The A can be represented as equation (1) states. U is an orthogonal matrix (m x m), V an orthogonal matrix (n x n) and S a diagonal matrix (m x n) with the singular values of A (Cao, 2006).

$$A = USV^T \tag{1}$$

The matrix U contains the singular vectors of A, which are arranged in order of *representativity* of the matrix decomposed. The first singular vector can be perceived as the stack of the information (Hopcroft, 2008), In our study case, this vector has the ability to represent the variations in local statistics of the image (Chandra, 2002).

The expansion of the matrix A is given by the columns of u_i and v_i of U and V respectively. And the singular values λ_i in S. The representation of each singular vector over A (see figure 2) is given by the square root of each singular value λ_i (Andrews & Patterson, 1976).

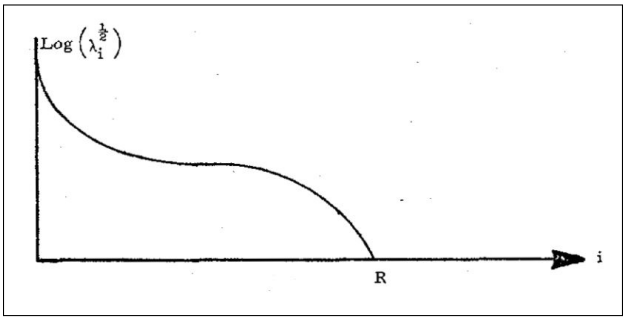


Figure 2: Schematic graph of the representation of each singular vector u_i monotonically ordered in logarithm scale. y axis represents how much a singular vector represents the data, on the other hand R is equal the number of columns m of A . Figure modified from (Andrews & Patterson, 1976).

3.b Principal Component Analysis *PCA*

PCA is a data reduction technique that is commonly used when dimensionality of a data-set needs to be reduced in a multi-dimensional space. This technique is broadly used for reducing the number of bands of multi-spectral imagery as it helps to find relevant level of correlation between the bands of the image (Jensen, 2005).

In order to reduce the dimensionality of a data-set with interrelated variables in terms of the variability of each one, a new set of variables is transformed (Principal Components PCs) so each one is uncorrelated (see figure 3). The PCs are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe 1, 2002). Suppose x_i is a vector of p random variables having maximum variance where i states variable measured in a given data-set. Each PC then would be a linear function $\alpha_i' x$ with length p so that:

$$\alpha_i^{'} x = \alpha_{11} x_1 + \alpha_{12} x_2 + ... + \alpha_p x_p = \sum_{j=1}^{p} \alpha_{1j} x_j \tag{2}$$

Then all of the *ith PCs* are representing a different amount of the variation. However, most of the variation in x will be accounted by a number c of PC where $c << p$. After imposing a normalisation over the data-set, it is possible to find the PCs in terms of the eigen vector and values of a matrix formed by the number of observations x (Jolliffe 1, 2002).

$$(\Sigma - \lambda I_p)\alpha_p = 0 \tag{3}$$

Where λ_p are the corresponding eigen values of the Matrix of observations and α_i its corresponding eiven vectors (PCs). To go further in the theory behind PCA refer to (Jolliffe 1, 2002).

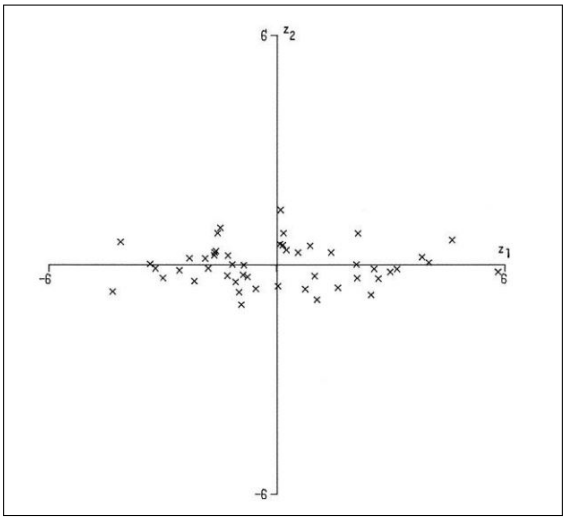


Figure 3: Set of observations $p = 50$ of a random variable x with respect to the first two PC (z_1 and z_2). The data displays that the variability of x is better represented by its first PC. Figure taken from (Jolliffe 1, 2002).

3.c Machine Learning - Random Forest

Machine Learning is a sub-discipline of data-science focused on mimic the process of human thinking for solving problems. In other words, it is a system that "learns" from experience based on different methods and parameters (Méndez Rojas, 2020).

A Random Forest classifier is a subset of individual learners (trees). Each one votes on a overall classification for a given set of inputs (see figure 4). A Random Forest classifier searches for the main characteristic from a sub-group of random characteristics (Méndez Rojas, 2020). Combining a large number of parallel classifiers the error and variance of the results decreases (Misra & Wu, 2020).

For a supervised classification of a image (ML model trained with labelled data-set that represents all the possible configurations of each class to be predicted). Each tree n is going to classify a pixel x in a category w or z. This classification is going to depend on a single characteristic $f(n)$ and a threshold θ_n. Depending on the value of $f(n)$, the pixel x is going to

be classified in one of the two classes (Méndez Rojas, 2020) and (De La Hoz, 2020):

$$x \in N^w \iff X_{f(n)} < \theta_n$$

$$x \in N^z \iff X_{f(n)} \geq \theta_n$$

The *ML* model implemented is a Random Forest classifier from the Python package scikit-image.

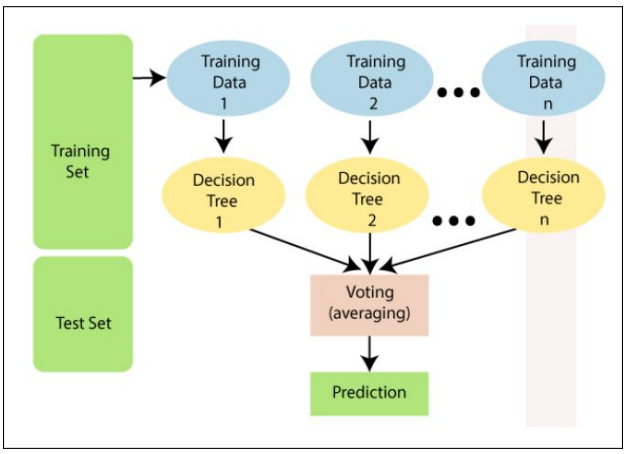


Figure 4: Random Forest model. Each *training data* column represents an individual classifier or tree. The prediction of each tree will be accounted at the end in order to produce a final prediction based on the result of each. Figure taken from (Méndez Rojas, 2020).

4 Procedure and Results

This section focuses on the processes made over the data-set and relevant parameters such as pre-processing and data parametrization.

4.a Procedure

The Section of the raster image used is composed of 704 pixels on x (Raster width) and 357 pixels on y (Raster height). Before applying the *SVD*, it is necessary to perform a scaling and centring all of the matrix begin processed (each band from table 1). The transformation of predictor features is often necessary and could vary depending on the data-set and the problem being solved. This is mostly done in order to improve the efficiency and reliability of the model. In our case, centering is made in order to remove the mean of the independent variable (mean of 0). While scaling, in a similar way, predictor variables are divided by their standard deviation. (standard deviation 1) (Jolliffe 2, 2002) and (Abdygaziev, 2020).

The result of the singular value decomposition is shown in figure 5. The first singular vector for each band is representing the stack of that band and therefore is the best representation of the matrix that forms it (Barrett & Beroza, 2014).

Each singular vector showed in figure 5 will form a column in a matrix U_{SVD}. Each band would represent a feature inter-correlated to each other. The process of *PCA* was performed by two methods. The first method was made through *numpy* implementation (following 1) Meaning and Centering the data. 2) Calculate the Covariance Matrix 3) Compute the Eigenvalues and Eigenvectors and 4) Sort Eigenvalues in descending order). The second method was made using

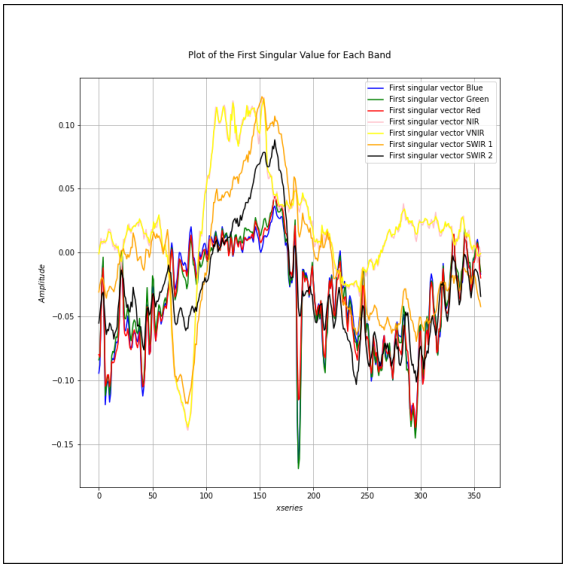


Figure 5: Each first singular vector of each band. The image shows the representation of each band in x which is the width of itself (in other words each row of data inside the matrix A). This image shows how each band behaves compared to the others in the same pixel, it recalls the spectral firm of each band over the image.

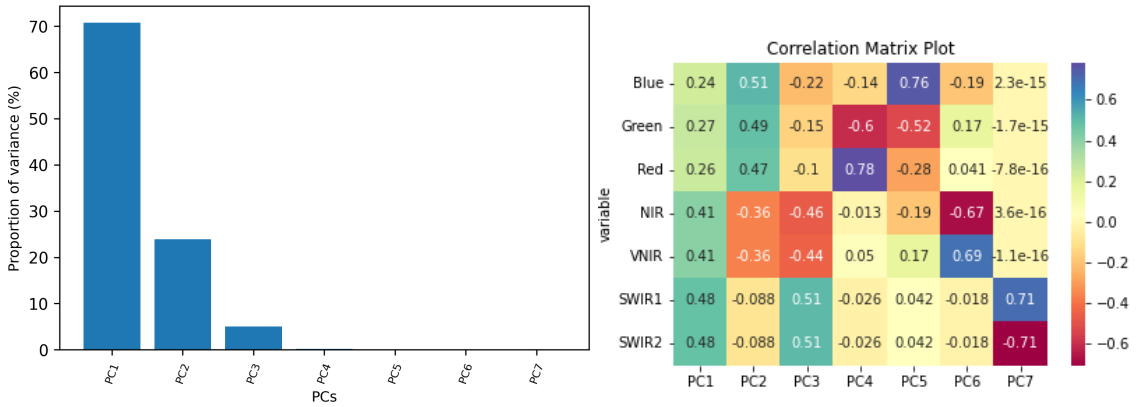
the *sklearn* package and the purpose of the second method was to validate the results obtained using *numpy*. The difference between the eigen values of the matrix U_{SVD} is showed in table 2.

numpy PCA	sklearn PCA	Difference
$1.08410016e^{-02}$	$1.08410016e^{-02}$	$6.93889390e^{-18}$
$3.66070820e^{-03}$	$3.66070820e^{-03}$	$2.60208521e^{-18}$
$7.72819795e^{-04}$	$7.72819795e^{-04}$	$1.08420217e^{-18}$
$3.00888250e^{-05}$	$3.00888250e^{-05}$	$3.86247024e^{-19}$
$7.60969854e^{-06}$	$7.60969854e^{-06}$	$1.01135734e^{-18}$
$3.84673275e^{-06}$	$3.84673275e^{-06}$	$1.46536700e^{-19}$
$4.23516474e^{-22}$	$4.23516474e^{-22}$	$4.23516474e^{-22}$

Table 2: Validation results from *numpy* and *sklearn* $PCAs$. Both methods where applied to validate the result obtained through *numpy*. The eigen values of the matrix U_{SVD} where used as metric in the comparison.

Once completed the PCA over the matrix formed by the stack of each band (SVD). The representation of the variability of the bands and its correlation with each PC is showed in figure 6. From this figure the selection of the 3 bands is assessed, and two bands where selected from PC_1 and the remaining one from the PC_2 representation. The selected bands are $B12$, $B11$ and $B2$ (see table 1 for reference), this selection was made by choosing the bands that where having the higher component loadings over the first two PCs. Figure 7 shows the RGB and PCA band compositions that are going to be tested in the Random Forest classifier.

Both images where under the same conditions (in terms of hyperparameters and data-set size) when performing the ML model. Each image was cropped into 10000 segments that could or not have a piece of labelled data. Three classes where used for this exercise 70% of the total was used to training and the remaining 30% for validation (note the simplicity of the problem



(a) Principal Component retention of original variables (*SVD* of each band).

(b) Correlation matrix between each band and the calculated *PCs*.

Figure 6: *a* shows the variability captured by each *PCs* calculated, PC_1 and PC_2 represents the most variability of the data-set (0.7 and 0.23 respectively). *b* displays the component loadings or weights (these represent the correlation between each band and each *PCs*).

and the small data-set used, as the purpose is to compare both images and not actually implementing a *ML* model at best conditions).

1. Class 1 - Water : Any body of water such as a river or a lake (15 samples). Segments for training are 3287 and 3587 for *PCA* and *RBG* respectively.
2. Class 2 - plants: Any land cover made of plant matter (putting together forest and grass cover) (15 samples). Segments for training are 1878 and 1705 for *PCA* and *RBG* respectively.
3. Class 3 - Land: Any land cover which is not water or matter plant (putting together natural soil and man-made elements) (15 samples). Segments for training are 1319 and 920 for *PCA* and *RBG* respectively.

The data-set, the images and the codes are in the same public repository named at the beginning of the document: Public Repository Data-set.

4.b Results

The overall result clearly shows classification of the image created by *PCA* and *SVD* outperforms the classification over the traditional *RGB* image. Before discussing the classification metrics for each image, it is necessary to compare both classifications in order understand the difference between the two images, see figure 8. The first thing to notice is that the classification over the *PCA* image delimiters the urban and green cover with better definition. Similarly, small types of coverage within the largest such as the lake at William Hawrelak Park Pavillion (Edmonton) or the green matter within the city (just like Main Quad at University of Alberta) are better defined and are having less mistakes in the classification.

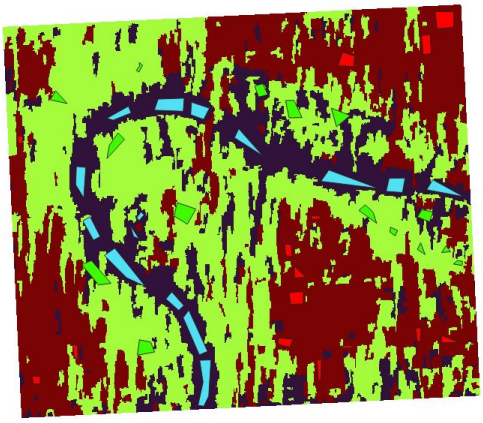
5 Analysis

The assessment of the model was made through the python package *sklearn.metrics*. The metrics used are explained in short using as reference (Melamed et al., 2003) and its values are shown in table 3:

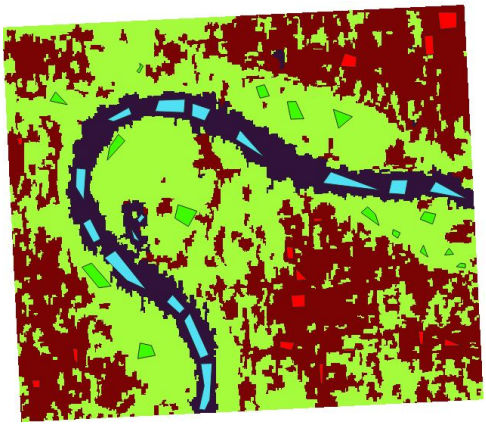
(a) RGB composition for base reference on performance of the Random Classifier (bands $B4$, $B3$ and $B2$)

(b) PCA composition using selected bands according to PCA results (bands $B12$, $B11$ and $B2$)

Figure 7: *a* shows RGB base composition that is going to be taken as reference in order to determine whether PCA analysis is successful or not. *b* new composition to be tested against RGB. Differences in land cover are clear when comparing both compositions, specially plant cover and water.



(a) Classification of the RGB image in three different classes.



(b) Classification of the PCA image in three different classes.

Figure 8: Right shows the result for the Random Forest classifier on the image build by PCA method. Left shows the same ML classifier on a RGB composition. The geometric shapes correspond to the training samples for each category: red for land, green stands for plants and blue for water.

1. Precision : This term accounts for the total of positive predictions for a class in comparison to all positive predictions of the model. In orther words, it compares a set of candidate items Y to a set of reference X, $\text{precision}(Y \mid X) = \frac{|X \cap Y|}{|Y|}$.
2. Recall : Similarly to precision, recall is the percentage of correct positive predictions from a relative point of reference considering all actual positives, $\text{recall}(Y \mid X) = \frac{|X \cap Y|}{|X|}$.
3. F1 score : "A weighted harmonic mean of precision and recall. The closer to 1, the better the model" (Zach, 2022).

Class and type of image	Precision	Recall	f1-score
Water PCA	1.00	0.99	0.99
Water RGB	0.76	0.99	0.86
plants PCA	0.97	1.00	0.98
plants RGB	0.99	0.72	0.83
land PCA	1.00	1.00	1.00
land RGB	0.94	1.00	0.97

Table 3: Classification report for both Random Forest classifiers. F1 score gives a better perspective of the behaviour of the model. Moreover, there are more ways of measure the performance of a model here just 3 parameters are being selected.

Table 3 shows that the PCA image was superior in terms of the positive predicted variables in contrast to the RGB implementation. This results suggest that the approach followed in study was positive as the metrics shows the potential of Feature Engineering for improving Machine Learning Methods.

6 Conclusions

This paper presents an approach to better select the information from remote sensing data-sets in order to implement Machine Learning models faster (less data to process), and in a more reliable way as the prepossessing of data (Feature Engineering) simplifies the identification of main characteristics for predictions in the model.

This study found that the decomposition of each band of the multi-spectral image can be decomposed using SVD and use this "stack" on different statistical approaches. PCA was selected on this study as the variability of a data-set such as this one represents no other thing than the different land cover types. Both RGB and PCA images where tested in a ML Random Forest Classifier in order to compare the results and test whether Feature Engineering is a key factor for ML or not.

The Image processed with Feature Engineering outperforms the classical RGB approach, suggesting that this methodology works. Therefore, the next step would be applying this methods to a stronger and more complete data-set. This should produce a better ML model without the necessity of using all bands of the image.

References

Abdygaziev, A. (2020) Data transformations: Centering amp; scaling, Medium. Medium. Available at: https://medium.com/@aabdygaziev/data-transformations-centering-scaling-7bd48a530595 (Accessed: November 07, 2022).

Anderson, M. R., Antenucci, D., Bittorf, V., Burgess, M., Cafarella, M. J., Kumar, A., ... Zhang, C. (2013, January). Brainwash: A Data System for Feature Engineering. In Cidr.

Andrews, H., Patterson, C. (1976). Singular value decompositions and digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(1), 26-53.

AskPython (2020) Principal component analysis from scratch in Python, AskPython. Available at: https://www.askpython.com/python/examples/principal-component-analysis (Accessed: November 05, 2022).

Barrett, S. A., Beroza, G. C. (2014). An empirical approach to subspace detection. Seismological Research Letters, 85(3), 594-600.

Bedre, R. (2021) Principal Component Analysis (PCA) and visualization using python (detailed guide with example), Data science blog. Available at: https://www.reneshbedre.com/blog/principal-component-analysis.html (Accessed: November 13, 2022).

Brownlee, J. (2019) How to calculate the SVD from scratch with python, Machine Learning Mastery. Available at: https://machinelearningmastery.com/singular-value-decomposition-for-machine-learning/ (Accessed: November 13, 2022).

Cao, L. (2006). Singular value decomposition applied to digital image processing. Division of Computing Studies, Arizona State University Polytechnic Campus, Mesa, Arizona State University polytechnic Campus, 1-15.

Chandra, D. S. (2002, August). Digital image watermarking using singular value decomposition. In The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002. (Vol. 3, pp. III-III). IEEE.

De La Hoz, E. C. (2019) Google Cloud Platform Big Data and Machine Learning Fundamentals. Coursera, FEB2ZXRXDAHA, course notes (delivered September 2019).

De La Hoz, E. C. (2022) Feature Engineering with Computational Methods for Machine Learning in Remote Sensing Applications, GEE sorce code (Version 1.0) [Source code]. https://code.earthengine.google.com/bcf6d6a35ffb3e1d226d3428525af842

De la Hoz Lozano, E. C. 2018) Subspace Detection de la Microsismicidad en el Mar de Mármara. Universidad de los Andes, Colombia. https://repositorio.uniandes.edu.co/

Hopcroft J. (2008) Computer Science in the Information Age. In: Preparata F.P., Wu X., Yin J. (eds) Frontiers in Algorithmics. FAW 2008. Lecture Notes in Computer Science, vol 5059. Springer, Berlin, Heidelberg

Jensen, J.R. 2005. Introductory digital image processing: A remote sensing perspective, 3rd Ed. Prentice Hall series in Geographic Information Science, Upper Saddle River, N.J., pp. 526.

Jolliffe, I.T. (2002) Principal component analysis. New York: Springer (Springer Series in Statistics (SSS).

Jolliffe, I. T. 2002. Principla Component Analysis. 2nd ed. Springer. (Springer Series in Statistics (SSS).

Melamed, I. D., Green, R., Turian, J. (2003). Precision and recall of machine translation. In Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers (pp. 61-63).

Méndez Rojas, J. (2020). Deforestación en la RNN Nukak, el PNN Chiribiquete y sus alrededores entre 1990 y 2020, utilizando algoritmos de Machine Learning y sus cálculos de precisión. Universidad de los Andes, Colombia. https://repositorio.uniandes.edu.co/

Misra, S., Wu, Y.. 2020. Machine Learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. Elsevier

Zach (2022) How to interpret the classification report in sklearn (with example), Statology. Available at: https://www.statology.org/sklearn-classification-report/ (Accessed: November 14, 2022).