Design of an investment portfolio through Machine Learning and Swarm Intelligence

Dr. Juan Fernando García-Mejía., Dr. Pedro Enrique Lizola-Margulis., Dr. Everardo Efrén Granda-Gutiérrez., Dra. Yenit Martínez Garduño, Dr. Allan Antonio Flores-Fuentes. Dra. Laura Leticia Laurent Martínez.

[a] Universidad Autónoma del Estado de México Centro Universitario UAEM Atlacomulco, Carretera Toluca Atlacomulco Km 60 Atlacomulco Mex. CP 50450 Méx. Mex. Email fgarciam@uaemex.mx
[b] Universidad Autónoma del Estado de México Facultad de Contaduría y Administración, Cerro de Coatepec Toluca CP 50130 Méx. Email plizolam@uaemex.mx
[c] Universidad Autónoma del Estado de México Centro Universitario UAEM Atlacomulco, Carretera Toluca Atlacomulco Km 60 Atlacomulco Mex. CP 50450 Méx. Email eegrandag@uaemex.mx
[d] Universidad Autónoma del Estado de México Facultad de Contaduría y Administración, Cerro de Coatepec Toluca CP 50130 Méx. Email ymartinezg@uaemex.mx
[e] Universidad Autónoma del Estado de México Centro Universitario UAEM Atlacomulco, Carretera Toluca Atlacomulco Km 60 Atlacomulco Mex. CP 50450 Méx. Email aafloresf@uaemex.mx
[e] Universidad Autónoma del Estado de México Facultad de Contaduría y Administración, Cerro de Coatepec Toluca CP 50130 Méx. Email llaurentm@uaemex.mx

Abstract

An investment portfolio is a collection of financial instruments where an investor allocates certain amounts of money to obtain a maximum profit with a minimum risk, modeled by Markowitz's Optimal Portfolio Theory, which is why the selection of financial assets and the allocation of amounts to invest are key. In this proposal, a single-objective optimization problem is proposed with the aim of maximizing the profit/risk ratio. The selection in the proposal documented in this paper is carried out by means of a clustering algorithm known as kmeans on the assets of S&P/BMV IPC of the Mexican Stock Exchange. On the other hand, the allocation of investment amounts is made after comparing the classical analytical method of the Generalized Reduced Gradient with two softcomputing methods: Particle Swarm Optimization, and the Whale Optimization Algorithm, a heuristic method that is based on whale hunting activities. Both methods are analyzed by means of statistics to determine the stability of both with respect to their executions. The results obtained in this work show that the best alternative studied is the Whale Optimization Algorithm, since it has a faster convergence and a higher objective function value in relation to the swarm of particles.

Keywords— Particle Swarm Optimization, and the Whale Optimization Algorithm, Markowitz's Optimal Portfolio Theory

1. INTRODUCTION

An investment portfolio is defined as a collection of financial instruments where an investor allocates certain amounts of money in order to obtain a maximum profit with a minimum risk. The financial instruments that make up an investment portfolio are public or government debt bonds, bank notes, corporate bonds, company ownership interests, investment funds, exchange-traded funds (ETFs), to name a few [1].

Currently, the importance of the design of investment portfolios lies in the fact that they can be constituted as an alternative to the low yields of Retirement Savings Systems, the volatility of the markets, as well as an adverse environment due to inflation [2].

In the construction of an investment portfolio, the Optimal Portfolio Theory is vital. Developed by the American economist Harry Markowitz, it establishes a theoretical scaffolding that aims to obtain maximum profit $\overline{r_p}$ definable in equation 1, with minimal risk R_p described in expression 2 with a constraint modeled in 3 [3]

$$\overline{r_p} = \overline{r_1} * \omega_1 + \overline{r_2} * \omega_2 + \overline{r_3} * \omega_3 + \cdots + \overline{r_n} * \omega_n \tag{1}$$

$$R_p = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_i * \omega_j * \sigma_{ij} \tag{2}$$

$$\sum_{i=1}^{n} \omega_i = 1 \tag{3}$$

where $\overline{r_p}$ represents the average expected return of the portfolio composed of $\overline{r_i}$ Returns on the assets of the elements that make it up for i $\geq 1,2,3 \ldots,$ n. On the other hand, ω_i It represents the amounts to be invested in each *i-th* asset, in addition to the equations described above σ_{ij} represents the covariance of *i-th* and j-th returns [3].

It should be noted that one of the fundamentals postulated by Markowitz is diversification, that is, that the investment amounts are calculated in terms of the variation in the returns of the assets that will be combined in the portfolio.

The selection of financial assets and the amounts to be invested in them is carried out through the analysis of the return and its coefficient β which is defined as the variability of a stock's performance r_i with respect to the average return r_m of the environment in which they are marketed, this can be represented mathematically by means of the expression 4

$$\beta = Cov(r_i, r_m)/s^2(r_m) \tag{4}$$

The selection of assets and investment amounts based on average returns and ratios can be made through the experience of financial specialists based on the investor's profile and degree of risk aversion [3]

An alternative for determining investment amounts ω_i It is by minimizing the expression 2 subject to the condition set forth in equation 3. This is done by means of a computational algorithm called Generalized Reduced Gradient (GRG) [1].

An alternative to GRG, according to a review of the specialized literature, is found in the application of artificial intelligence techniques, which is defined as a set of algorithms that aim to emulate thought and learning processes. A category of techniques that are supported, specifically on biological principles, is softcomputing, which is applied in the aesthetic solution of optimization problems which can be combinatorial, which consists of finding a $\vec{x}$ / $f(\vec{x}) = \max f\ \forall\ \vec{x} \in\ \mathbb{Z}$ .The numerical in such a way that $\vec{x}$ / $f(\vec{x}) = \max f\ \forall\ \vec{x} \in\ \mathbb{R}$ [2].

In a review of the specialized literature, under the context of combinatorial optimization, it is shown that the maximization of profits and minimization of investment risk can be carried out by means of Genetic Algorithms, attributed to John Holland theoretically supported by Mendel's Laws of Heredity and the Theory of Evolution of Species. Another solution approach is found in the combination of Genetic Algorithms, with the use of Artificial Neural Networks (NNA), proposed for the first time by Warren McCulloch and the mathematician Walter Pitts, with the biological scaffolding of Santiago Ramon y Cajal. Finally, the optimization of investment portfolios has been documented by means of Differential Evolution algorithms developed by Kenneth Price and Rainer Storn as a variant of the Genetic algorithms [3] [4].

An alternative to [ref] is found in the portfolio solution considering it as a numerical optimization problem, this is shown in [5] where a comparative study of the response of a Genetic Algorithm, a Differential Evolution and a Clonal Selection Algorithm is carried out, a heuristic supported in the functioning of the immune system of higher mammals.

In 1997, David Wolpert and William Macready developed a theoretical proposal that determines that there is no optimization algorithm that is characterized by its universality, that is, that it can solve all optimization problems in all contexts. This allows you to follow a line of research whose object of study is the application and development of computational techniques for the field of finance. Based on this, this proposal proposes the selection of financial assets by means of a clustering algorithm to select financial assets and the calculation of investment amounts by means of techniques called swarm intelligence [6].

2. THEORETICAL SUSTENANCE

As mentioned in the previous section, the selection of financial assets is done by a financial specialist, which inherently implies a bias. An alternative to this is found in the selection of artificial intelligence techniques, a branch of artificial intelligence that aims to imitate human behavior through learning processes and knowledge acquisition.

A machine learning technique par excellence is clustering, which has applicability in data mining, since it is used to divide a set of unlabeled D data with vector representation into $\vec{x} \in \mathbb{R}^N K$ sets, therefore it is considered an unsupervised learning algorithm. The most common is the *K-means* cluster, which has the functionality of generating datasets in such a way as to minimize the sum of the distance between the data and its assigned set. This is shown in algorithm 1[7]

It is important to note that the selection of the optimal cluster number can be estimated by the Elbow method, which uses the calculation of within cluster sums of squares determined by equation 5

Algorithm 1 *K-means* clustering algorithm

```
Requiere: K, number of clusters; D, a data set of N points
Ensure: A set of K clusters
    Initialization
    repeat
        for each point p in D do
            find the nearest center and assign p to the
            corresponding cluster
        end for
        update clusters by calculating new centers using
        mean of the members
    until stop-iteration criteria satisfied
    return clustering result
```

$$wcss = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad (5)$$

Where μ_k represents the K-th centroid and x_i el i-th D data set. The method allows you to select a value of K in terms of the transition of the values of $wcss$.

The other important stage in the construction of an investment portfolio is the calculation of the amounts to be invested in each of the assets, which according to what has been shown in the state of the art can be done by means of evolutionary algorithms.

An alternative to the evolutionary algorithms theoretically supported by Charles' theories. Darwin and Gregory Mendel's laws of inheritance are the algorithms of swarm intelligence, which emulate the collective behavior of some decentralized systems such as schools and flocks of birds, to mention a few examples. The pioneers in this area are Gerardo Beni and Jing Wang in the 1990s [8]. The original swarm intelligence algorithm is Particle Swarm Optimization (PSO) by James Kennedy and Russell Eberhart.

The PSO has its theoretical scaffolding in social behavior, of the cooperative type, which presents flocks of birds or schools of fish, when they fly to a certain point or perform a search for food. In this type of collective intelligence, the movement (characterized by direction, velocity, and acceleration) of the individuals belonging to the collective is the result of combining the individual decisions of each of the behaviors of the others [9].

In the PSO, a population of i particles represented by $\vec{x}_i = \{x_{i,1}, x_{i,2}, x_{i,3}, \ldots, x_{i,d}\}$ which are solutions to an optimization problem $f(\vec{x}_i)$ of dimension *d*, each *i-th* particle of the swarm flies over the solution space to new positions $\vec{x}_i$, with a velocity vector $\vec{v}_i = \{v_{i,1}, v_{i,2}, v_{i,3}, \ldots, v_{i,d}\}$ with an acceleration parameter *w*, During the development of the algorithm, an optimal generation must be found p_g and a local optimum definable by means of p_i. This is representable by algorithm 2 [9].

Algorithm 2 Particule Swarm Optimization

Initialize population
for t= 1: maximum generation
  for i= 1: population size **do**
    if $f\left(x_{i,d}(t)\right) < f\left(p_i(t)\right)$ **then** $p_i(t) = x_{i,d}(t)$
      $f\left(p_g(t)\right) = \min_t \left(f\left(p_i(t)\right)\right)$
    end
    for d = 1: dimension
      $v_{i,d}(t+1) = wv_{i,d}(t) + c_1r_1\left(p_i - x_{i,d}(t)\right) + c_2r_2\left(p_g - x_{i,d}(t)\right)$
      $x_{i,d}(t+1) = x_{i,d}(t) + v_{i,d}(t+1)$
      if $v_{i,d}(t+1) > v_{max}$ **then** $v_{i,d}(t+1) = v_{max}$
      else if $v_{i,d}(t+1) < v_{min}$ then $v_{i,d}(t+1) = v_{min}$
      end
      if $x_{i,d}(t+1) > x_{max}$ **then** $x_{i,d}(t+1) = x_{max}$
      else if $x_{i,d}(t+1) < x_{min}$ **then** $x_{i,d}(t+1) = x_{min}$
      end
    end
  end
end

After the development of the PSO, other heuristic methods have emerged that aim to improve the convergence of the search for an optimum, one of these alternatives is The Whale Optimization Algorithm (WOA).

The WOA is a heuristic that allows you to solve optimization problems of the type that is inspired by the hunting behavior of humpback whales in their natural environment. They are considered the largest mammals; adults can measure up to 30 meters and weigh 180 tons [10].

An interesting aspect about the behavior of humpback whales is the behavior they exhibit to obtain their food, which is known as the bubble net search method, and they use it to hunt schools of krill or small fish near the surface, looking for them by creating bubbles along a circle or number 9. It should be noted that this behavior can be modeled by means of mathematics, as shown in expression 6 [10].

$$\vec{D} = \vec{C}X^* - \vec{X}(i) \tag{6}$$

$$\vec{X}(t+1) = \vec{X}^* - \vec{A}\vec{D} \tag{7}$$

Where $\vec{C} = 2r$, $A = 2ar - a$, $\vec{X}^*$ It is the best vector of the dam and $\vec{X}$ indicates the vector that represents the best position of the whale [11].

$$X(t+1) = D'e^{bl}cos2\pi l + X^* \tag{8}$$

Where *b* It's a constant and *l* a random number between -1 and 1. It is important to consider that whales, after locating their prey, swim around them in smaller and smaller circles, assuming that there is a *p* probability of 50% to choose an enveloping mechanism of contraction or the spiral model, to update the position of the whales this is described by means of the expression 9.

$$X(i+1) = \begin{cases} X^* - AD\ ;\ p < 0.5 \\ D'e^{bl}cos2\pi l + X^*;\ p >= 0.5 \end{cases} \tag{9}$$

It should be noted that the computational model, shown in algorithm 3 was developed by Mohammad H. Nadimi-Shahraki in 2016 [11].

Algorithm 3 Whale Optimization Algorithm

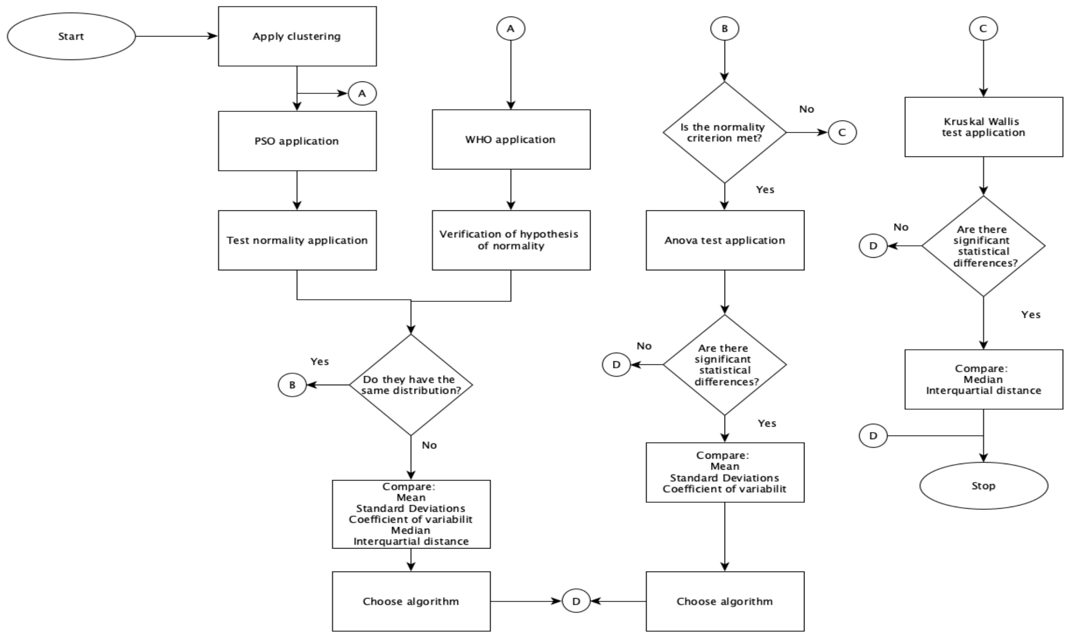
Initialize population
Calculate fitness $\boldsymbol{f(\vec{x}_i)}$
X^*= the best agent search
While i < maximum number of iteration **do**
  for every solution **do**
    $a = 2 - t\dfrac{2}{max\ iter}$
    $A = 2ar - a$
    $C = 2r$
    $p = rand[0,1]$
    $l = rand[-1,1]$
    $D = CX^* - X(i)$
    $D' = |X^* - X(i)|$
      if $(p < 0.5)$ **then**
        if $(|A| < 1)$ **then**
          $X(i+1) = X^* - AD$
        else if $(|A| > 1)$
          X_{rand} *solution is generated*
          $X(i+1) = X_{rand} - AD$
        end
      else if $(p >= 0.5)$
          $X(i+1) = D'e^{bl}cos2\pi l + X^*$
      end
  end
  Compute the fitness of every solution
  Update X^*there is a better solution
  $i = i + 1$
end

3. METHODOLOGY

For this work, the 35 financial assets of the Mexican Stock Exchange are considered as objects of study, specifically the S&P/BMV IPC index, which consists of 36 assets, in which the performance and its average beta from June 2017 to June 2023 are considered.

The methodological scheme to be addressed in this paper is shown in Fig 1, in which it can be observed that a clustering technique is performed, specifically the cluster by *k means* to select the assets that will constitute the proposed portfolio.

Fig. 1. Methodological scheme



Once the financial assets have been obtained, two e-optimization algorithms are executed, first one by the particle swarm algorithm (PSO) and then by means of the WOA whale optimization algorithm.

In order to determine if there are statistically significant differences between the algorithms, a statistical study is carried out which consists first of all in the approval/rejection of the assumption of normality, this is carried out by means of a Kolmogorov test, once the assumption of normality is determined, an Anova test is applied, in the case that the assumption of normality is verified for both algorithms. or the Kruskal Walls test if for both algorithms this assumption is rejected. Once the behavior of both algorithms has been determined, a comparison of statistical parameters is carried out, to determine which is the best algorithm, both in terms of profit/risk ratio as well as convergence.

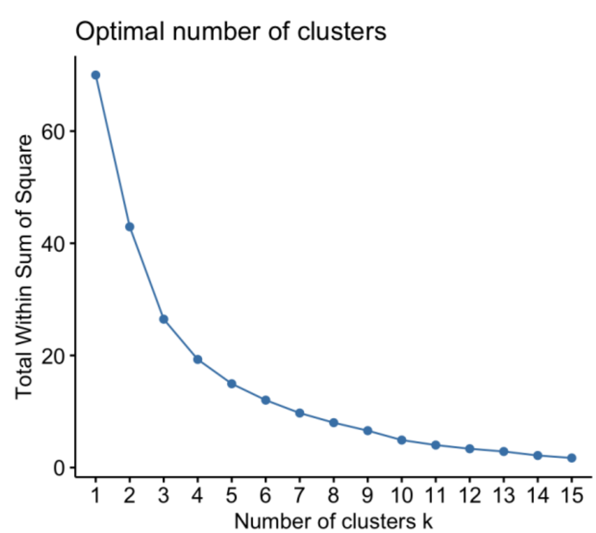
The fitness function that is used is the relationship between profit and risk, so this can be shown in expression 10.

$$f_{obj} = \frac{\sum_{n=1}^{N} \overline{r_n} * w_n}{\sum_{n}^{N} \sum_{m}^{N} w_n * w_m * \sigma_{nm}} \tag{10}$$

4. RESULTS

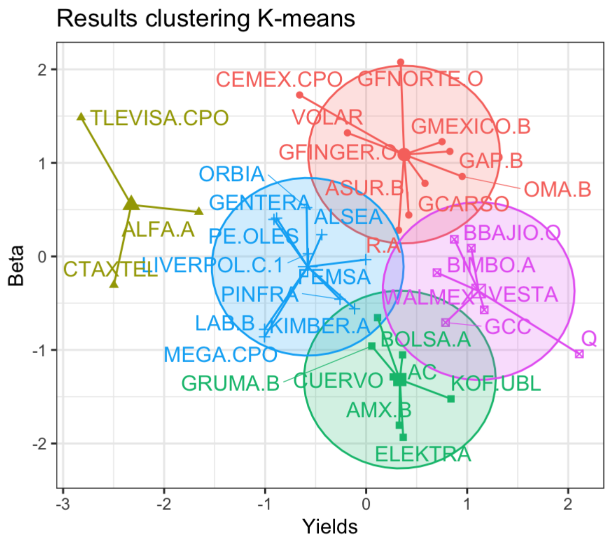
As shown in Figure 1, the first step in this methodological process is the selection of assets through the application of clustering, specifically the k-means method. As a first step, it is necessary to determine the number of clusters by means of the inflection of the curve shown in Figure 2, this by means of a technique called the elbow method, which consists of determining the optimal number of clusters, which is reflected in a minimum distortion, that is, the space between elements and the centroid of the cluster. By applying the k-means algorithm with a k=5 as shown in Figure 2, the groups shown in Figure 3 are generated

Fig. 2. Elbow Chart



Source: Authors' own creation

Fig. 3. Clusters obtained



Source: Authors' own creation

In Figure 3 it is possible to observe the clusters formed by the method of k means, for k=5 as shown by the decay of the elbow in Fig 2. Clusters are grouped by their performance and their beta parameter, so this allows financial assets with negative returns and positive betas to be discarded.

From Fig 3 it is possible to construct the portfolio shown in Table 1, which was constructed by taking the assets with the highest return in each group, except for the grouping that has the negative returns.

Table 1. Assets of the proposed portfolio

Active	Yield	Beta
Q	8.819569e-04	0.5055619
Vesta	5.162075e-04	0.6640651
Femsa.UBD	6.028173e-05	0.8448731
Kimber.A	1.747484e-05	0.6679033
KOF.UBL	3.883542e-04	0.3437087
ELEKTRA	2.045508e-04	0.2052298
OMA	4.314582e-04	1.1442427
GAB.B	3.840923e-04	1.2344836

Subsequently, in accordance with what was indicated in the methodology section, a statistical study was carried out on the values of 40 executions of the proposed algorithms, giving a p less than 0.05, which indicates that the assumption of normality for both algorithms is rejected, therefore it is necessary to apply a non-parametric test. Originally, the Wilcox test was proposed obtaining a significance value p=2.2e-16, therefore it is considered that there are significant differences between both algorithms

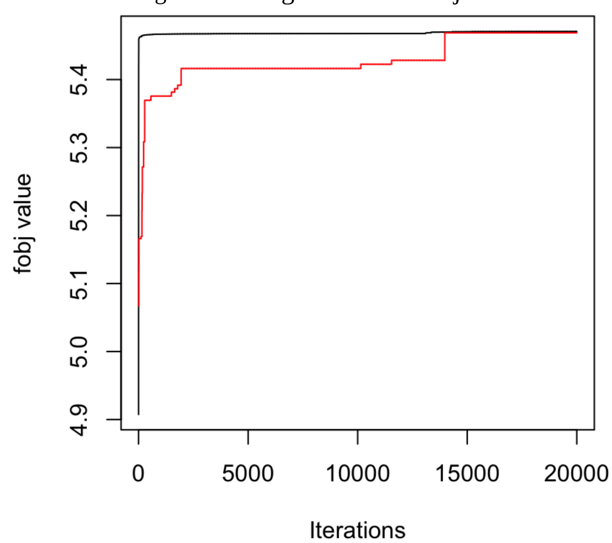
Therefore, Table 2 shows the results of the value obtained from the target function. It should be noted that the GRG-based solution method does not diversify the portfolio as it suggests investing all capital in a single asset. This table shows the amounts to be invested in each of the assets, in the first column are the amounts obtained by the WOA while the second corresponds to PSO.

Table 2. Assets of the proposed portfolio

Active	Amount with WOA	Amount to invest with PSO
Q	0.30033995	0.28956255
Vesta	0.21310850	0.22031294
Femsa.UBD	0.03271321	0.05782181
Kimber.A	0.02455626	0.04795786
KOF.UBL	0.10991190	0.11072677
ELEKTRA	0.26218006	0.23135580
OMA	0.04058151	0.01804613
GAB.B	0.01660862	0.02421615

Finally, it is possible to observe that Fig 4 shows the convergence of these algorithms, where the answer with the best convergence is the one obtained by means of WOA.

Fig. 4. Convergence of the fobj value



Source: Authors' own creation.

3. CONCLUSIONS

It is necessary to understand that heuristic algorithms are a viable alternative to traditional methods. Based on this postulate, some questions can be formulated that allow us to expand this research:

1. How will another Swarm Intelligence-based heuristic work?
2. Is there a difference between evolutionary algorithms and algorithms based on swarm intelligence?

The k-means clustering method is an alternative to asset selection, but it is necessary to test the effects of other algorithms such as those based on fuzzy logic. Also, as a means of selection, it is necessary to test hierarchical groupings, to constitute an alternative for the selection of assets.

The use of evolutionary algorithm techniques and, above all, swarm intelligence proves to be an alternative solution, which has only been explored in evolutionary algorithms inspired by the laws of inheritance, but according to the systematized review of the literature, numerical problems in investment portfolios have not been treated by means of swarm intelligence techniques such as the one shown in this work.

Based on the above, it is proposed as a future work to use various artificial intelligences based on the concept of swarms, such as shoals of fish, to mention a few.

4. REFERENCES

[1] E. Kievskaya, "Analysis of modern approaches to the formation of the portfolio investor shares stock", Technol. Aud. Prod. Reserv., vol. 5, núm. 2(43), pp. 4–9, 2018.

[2] Y.-F. Chen, T. C. Chiang, y F.-L. Lin, "Inflation, equity market volatility, and bond prices: Evidence from G7 countries", Risks, vol. 11, núm. 11, p. 191, 2023.

[3] H. A. Sukardi, M. Sari, N. Nugraha, y I. Purnamasari, "Markowitz portfolio optimization applied on companies listed on the Indonesia Stock Exchange LQ-45", Ekono Insentif, vol. 17, núm. 1, pp. 34–48, 2023.

[4] P. A. H. Manurung, "Stock selection using semi-variance and Beta to construct portfolio and effect macro-variable on portfolio return", Turk. J. Comput. Math. Educ. (TURCOMAT), vol. 15, núm. 1, pp. 14–25, 2024.

[5] J. F. García-Mejía, E. Rodríguez-Lopez, Y. Martínez-Garduño, P. E. Lizola-Margolis, E. E. Granda-Gutiérrez, y F. E. Valdés-Medina, "Analysis of canonical heuristic methods for the optimization of an investment portfolio", en Handbook on Decision Making, Cham: Springer International Publishing, 2023, pp. 51–69.

[6] D. H. Wolpert y W. G. Macready, "No free lunch theorems for optimization", IEEE Trans. Evol. Comput., vol. 1, núm. 1, pp. 67–82, 1997.

[7] M. Ahmed, R. Seraj, y S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation", Electronics (Basel), vol. 9, núm. 8, p. 1295, 2020.

[8] G. Beni y J. Wang, "Swarm intelligence in cellular robotic systems", en Robots and Biological Systems: Towards a New Bionics?, Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 703–712.

[9] J. Kennedy y R. Eberhart, "Particle swarm optimization", en Proceedings of ICNN'95 - International Conference on Neural Networks, 2002.

[10] M. H. Nadimi-Shahraki, H. Zamani, Z. Asghari Varzaneh, y S. Mirjalili, "A systematic review of the whale optimization algorithm: Theoretical foundation, improvements, and hybridizations", Arch. Comput. Methods Eng., vol. 30, núm. 7, pp. 4113–4159, 2023.

[11] Nadimi-Shahraki, M., Zamani, H., Asghari Varzaneh, Z. *et al.* A Systematic Review of the Whale Optimization Algorithm: Theoretical Foundation, Improvements, and Hybridizations. *Arch Computat Methods Eng* **30**, 4113–4159 (2023). https://doi.org/10.1007/s11831-023-09928-7