https://doi.org/10.51483/IJMRE.1.4.2021.24-47 ISSN: 2710-141X

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Management Research and Economics

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Unsupervised Learning Diversification Applied on the Tunisian Stock Market Before and During the Covid-19 Crisis

Ahmed Rebai[1*], Louay Boukhris[2], Lotfi Ncib[3] and Mohamed Anis Ben Lasmer[4]

[1]Value Digital Services, Tunis, Tunisia. E-mail: ahmed.rebai@value.com.tn
[2]Value Digital Services, Tunis, Tunisia. E-mail: louay.boukhris@esprit.tn
[3]ESPRIT School of Engineering, Tunis, Tunisia. E-mail: lotfi.ncib@esprit.tn
[4]ESPRIT School of Engineering, Tunis, Tunisia. E-mail: mohamedanis.benlasmar@esprit.tn

Article Info

Volume 1, Issue 4, October 2021
Received : 21 June 2021
Accepted : 15 September 2021
Published : 05 OCtober 2021
doi: 10.51483/IJMRE.1.4.2021.24-47

Abstract

Financial data, related to companies listed on the Tunisian Stock Exchange, were collected and analyzed according to the methodology applied in machine learning on over two different time periods. A particular interest was focused on the periods before and during the Covid-19 crisis. The results obtained in this paper show, on the one hand, that an empirical diversification based on unsupervised learning algorithms is possible and on the other hand, a good coherence with the corporates financial state in Tunisia. This paper shows, for instance, that the k-means algorithm makes it possible to segment companies according to several criteria and to discover the aberrant behavior of certain companies with an abnormal financial situation. These results were confirmed by other outlier detection algorithms.

Keywords: *Unsupervised learning, Stock market, Finance, CAPM (Capital Asset Pricing Model), Machine learning, Asset management*

1. Introduction

Finance focuses on the decision to invest the resources available over a fixed time period in an unpredictable way in order to make a profit. Such a decision could therefore entail risks to the allocated resources. Minimizing these risks is then necessary, hence the need for diversifiable risk diversification as there is another incompressible factor that constitutes systemic risk (non-diversifiable). Traditionally, the CAPM (Capital Asset Pricing Model) theory historically offers a sound theoretical basis for achieving such diversification and obtaining portfolios of assets that reduce risk and maximize return through quadratic optimization. This theory is discussed at length in the papers by Markowitz (1952) and William Sharpe (Perold, 2004). Unfortunately, in terms of risk limitation, a portfolio maximizing the Sharpe ratio is not always the optimal portfolio, especially during very rare four sigma events, such as financial or health crises. This behavior was observed during the 2008 subprime crisis (Kaizoji and Miyano, 2018) and during the current Covid-19 crisis, mainly in March 2020 (Baker *et al*., 2020a). Faced with this limitation of classical theories, professionals (in asset management) alternate between mathematical techniques ranging from Monte Carlo simulations to hedging using options with the Black and Scholes (1973) model and empirical techniques such as Constant Proportion Portfolio Insurance (CPPI) (Cont and Tankov, 2007).

* *Corresponding author: Ahmed Rebai, , Value Digital Services, Tunis, Tunisia. E-mail: ahmed.rebai@value.com.tn*

Thanks to the numerous success of machine learning in several domains, nowadays its use in the field of finance is becoming widespread (Leung *et al*., 2021). Artificial Intelligence with its unsupervised learning techniques, offers algorithms that ease the detection of weak signals in large volumes of data. Indeed, the problem of diversification of a portfolio of financial assets in a parameterized space can be reformulated as a segmentation problem in the same space.

2. Methodology Followed During the Study

Nowadays, data science is gradually moving from the field of scientific research to the field of computer engineering. On the other hand, the use of data science in a managerial and financial context is increasing, the need to create a business-oriented strategy is important to limit the wide choice of existing models and consequently to understand and meet the needs of customers. Data scientists are now increasingly required to deploy these mathematical and algorithmic models quickly into production. This requires following an agile methodological approach to achieve the clients' objectives in a reasonable time. In contrast to classical working methods based on a pre-planned and sequential approach, a methodology based on a "problem-solving" approach seems indispensable. The integration of these requirements in the development cycle led us to follow one of two methodologies developed by the IBM company and which are known as "CRISP-DM" (Shearer, 2000) (Cross Industry Standard Process) and "IBM MasterPlan" (John, 2015). The IBM Master Plan methodology was the most suitable for our project as it includes a feedback stage, specific to the stock market experts we were in contact with. This methodology consists of the following nine points:

- Business understanding: As a team of data scientists and mathematicians we started by improving our knowledge in the field of market finance in general and asset management in particular. Then, call this step the understanding of the underlying business domain. During this step, we had continuous contact with financial analysts to define the problems, the project objectives, and the requirements of the solution from an asset management point of view.
- The analytical approach: We expressed the problem in its technical context (descriptive or inferential statistics, machine or deep learning, unsupervised learning or supervised learning...). An unsupervised learning combining clustering and anomaly detection was chosen as the analytical approach.
- Understanding the data: Descriptive statistics and visualization techniques were used to understand the content of the data, to assess the quality of the data. This step helps to find out whether additional data is needed or not from the client.
- The data preparation stage: As expected, this stage was the longest in our study and cost us 80% of the total time allocated. Features engineering, features selection, and dimension reduction are processes that are eventually found in the data preparation activities, including data cleaning (handling missing values, correct formatting), combining data from several sources (excel, html and csv files), and transforming data into more useful variables that are discussed in section 7. It is important to know that the more correctly the data is prepared, the easier it is for the models used during the modeling.
- The modeling stage: Given that our study is a diversification problem combining clustering and anomaly detection, we applied algorithms and models from unsupervised parametric learning such as K-means, DBSCAN, hierarchical clustering, and affinity propagation.
- Evaluation: During modeling and before deployment, it is essential to evaluate the model to understand its quality and to ensure that it correctly and fully addresses the business problem. The evaluation of the model involves the use of several metrics to interpret the quality of the model and its effectiveness in solving the problem. In section 7, we propose to explain the metrics used to evaluate our models.
- Deployment: We used a software architecture based on a Model-View-Template (MVT) specific to rapid prototyping using web technologies provided in Python like the micro frameworks dashplotly or flask.
- The feedback stage: We can consider this stage as the real evaluation of our study. Indeed, it is necessary to collect new data to confront the results of the model obtained with the actual with the new ones.

3. Management of a Portfolio of Financial Assets on the Tunisian Stock Exchange

Finance is the application of different economic principles to make decisions that involves the allocation of money under uncertain conditions. It provides the framework for making decisions about how to get money and what we should do with it once we have it. It is the financial system that provides the platform through which funds are transferred from the entities that have funds to the entities that need those funds (Maurizio *et al*., 2018). Finance is generally divided into two main branches:

- **Corporate Finance:** Corporate finance studies the investment and financing choices of companies in a certain world.
- **Market Finance:** Market Finance's main course is the study of financial markets. It includes among other things the variation in the prices of financial assets and portfolio management. Market finance is the sector of finance that is interested in the functioning and operations of financial markets under uncertainty.

A financial market is a physical or virtual place where the different market participants (sellers and buyers) trade financial products. Financial markets help finance the economy on the one hand and enable investors to make investments on the other. The financial markets can be classified into four major markets, depending on the nature of the assets traded, explained into the following points:

- **Stocks or shares:** A share represents part of the capital of a certain company. Holding a share means owning a part of the company and therefore having the right to receive dividends if the company distributes them, and possibly to participate and vote at general meetings of shareholders.
- **Dividend:** A dividend is a remuneration for the risk that the shareholder has taken by investing in a certain-listed company. It is derived from the profit made by a company and therefore varies according to the performance of that company during the concerned period.
- **Stock market index:** A stock market index is an important indicator that determines the performance of a market. Composed of a group of stocks, it represents a market, a sector of activity, or an economy. It shows the trend of the economy and the largest companies in the index. In Tunisia, there are two stock market indices which are Tunindex, which represents all listed Tunisian companies, and Tunindex 20, which represents the twenty most liquid listed companies with the largest market capitalizations.
- **Rate of return:** The return on a financial asset is the profit or loss on an investment over a certain period of time. A positive return indicates a profit while a negative return indicates a loss.

3.1. Diversification

Diversification is a fascinating concept in financial risk management that is used by various stakeholders such as investment managers, traders, quantitative analysts, and risk managers in monitoring financial market dynamics and building portfolios. First of all, after introducing diversification and explaining its important role, we will talk about four approaches used to achieve optimal diversification: the scientific approach (Markowitz, 1952; Fama and French, 1993 and 1996), the empirical approach or naive approach (DeMiguel *et al*., 2007), the new diversification strategies (Cont and Tankov, 2007) appearing after 2008 crisis, and the machine learning diversification strategies (L´opez de Prado, Marcos, 2019). A comparison between these four approaches will be given at the end of this section. Diversification consists in finding the best assets with their best contributions to build an efficient portfolio that allows reaping the highest premium corresponding to the risks incurred. In this respect, diversification is not an objective in itself, in other words, a rational investor should not be happy to hold a diversified portfolio because his ultimate goal is to reap the rewards. Indeed, suppose in an ideal world where this investor could predict the most profitable asset in the future then the problem is solved since he would have to invest all his wealth in this asset.

3.1.1. Scientific Diversification

This methodology is based on a rigorous mathematical formalism initiated by Markowitz in his famous 1952 article (1) and known as the modern theory of the CAPM. Indeed, the reward per unit of risk is defined in terms of the net ratio. The high ratio of a portfolio is the average return of the portfolio minus the risk-free. This ratio is called the Sharpe ratio defined by the following formula:

$$SR = \frac{(r_p - r_f)}{\sigma_p} \quad \text{...(1)}$$

where :

r_p is the return of the portfolio.

r_f is the risk free rate.

σ_p is the standard deviation of portfolio's excess return.

There exist multiple techniques to construct a portfolio using scientific diversification we can mention: The CAPM is a fundamental building block for many other asset pricing models in finance. This model is used in finance to determine the expected profitability of a risky asset, including equities. It was developed by William Sharpe and famous

economists. It starts from some simplified hypotheses (no transaction costs or taxes, all investors have identical perceptions regarding expected returns, short selling or buying a security does not affect its price, the market is completely free and all assets can be exchanged. . .). Unfortunately, these conditions cannot be applied in the Tunisian market. The CAPM is given by the following formula:

$$E[r_p] - r_f = \beta_p (e[r_m] - r_F) \quad \ldots(2)$$

where :

$E[r_p] - r_f$ is the excess expected return of a stock or a portfolio *P*.

$\beta_p(E[r_m] - r_F)$ is the excess expected return of the broad market portfolio β.

RF is the regional risk free-rate.

β_p is the portfolio beta, or exposure, to the broad market portfolio β.

In order to improve the CAPM model, which is a one-factor model, there is the FAMA-French model (three-factor model).

$$\begin{aligned} & E[r_i] - r_f \\ & = \beta_{i,MKT} \ E[r_m - r_F] \\ & + \beta_{i,SMB} \ E[SMB] \\ & + \beta_{i,HMS} \ E[HMS] \end{aligned} \quad \ldots(3)$$

The variables are explained as follows:

$\beta_{i,MKT}$ is the same β as the CAPM.

SMB means small minus big size stocks.

HML means High (B/P ratio) Minus Low (B/P ratio) stocks.

Carhart four-factor model (Carhart, 1997) further improves the FAMA-French model. As follows the Carhart model in it's regression form.

$$\begin{aligned} & r_i - r_f \\ & = \alpha + \beta_1 (r_m - r_f) + \beta_{2i} (SMB) \\ & + \beta_{3i} (HML) + \beta_{4i} (WML) + \varepsilon_i \end{aligned} \quad \ldots(4)$$

The variables are explained as follows:

r_i the return on asset *i*.

r_f the risk-free interest rate in government bonds.

α intercept of the regression line.

r_m return of the market portfolio.

(SMB) return of size factor.

(HML) return of the BE/ME factor.

(WML) return of the momentum factor.

$\in_i$ residuals of the regression model.

$\beta_{1-2-3-4}$ beta values of the independent variables.

In order to make a proper comparison between scientific and machine learning diversification, we applied the CAPM theory to the Tunisian market during two periods. Thus, the data used was scraped from the investing.com site according to the following two periods:

- A period before the crisis of Covid-19: March 2019-March 2020.
- A period during the crisis of Covid-19: March 2020-March 2021.

The results of the application of the CAPM model are presented in Figures 1, 2, 3, and 4. In the Annual Return vs. Volatility Plan, the efficient frontier is constructed using a simulation of 100,000 portfolios composed of companies listed on the Tunisian Stock Exchange with a positive Sharpe ratio. The red star represents the Maximum Sharpe Ratio (MSR)

	sharpe	ret	stdev	Mpbs	Tuninvest	Soc. Immob. Tuniso-Seoud.	Tawasol Group Holding SA	Hexabyte	Soc Ind. dapp et de mat. elec	Magazin Gneral	...
56789	4.006717	0.399045	0.099594	0.040462	0.039946	0.039399	0.038651	0.036395	0.035678	0.034008	...

1 rows × 47 columns

	sharpe	ret	stdev	El Wifack Leasing	Magazin Gneral	Euro-Cycles	Sotumag	Tuninvest	Attijari Leasing	Cerealis Sa	...
403	3.848936	0.319856	0.083102	0.039151	0.038258	0.038104	0.037561	0.034367	0.033809	0.032357	...

1 rows × 47 columns

Figure 1: The Maximum Sharp Ratio is Represented in the Upper Row Where We Expose the Sharp Ratio, the Return, the Standard Deviation and the Weights of the Stocks with Positive Returns. The Lower Row Shows the Minimum Variance Portfolio

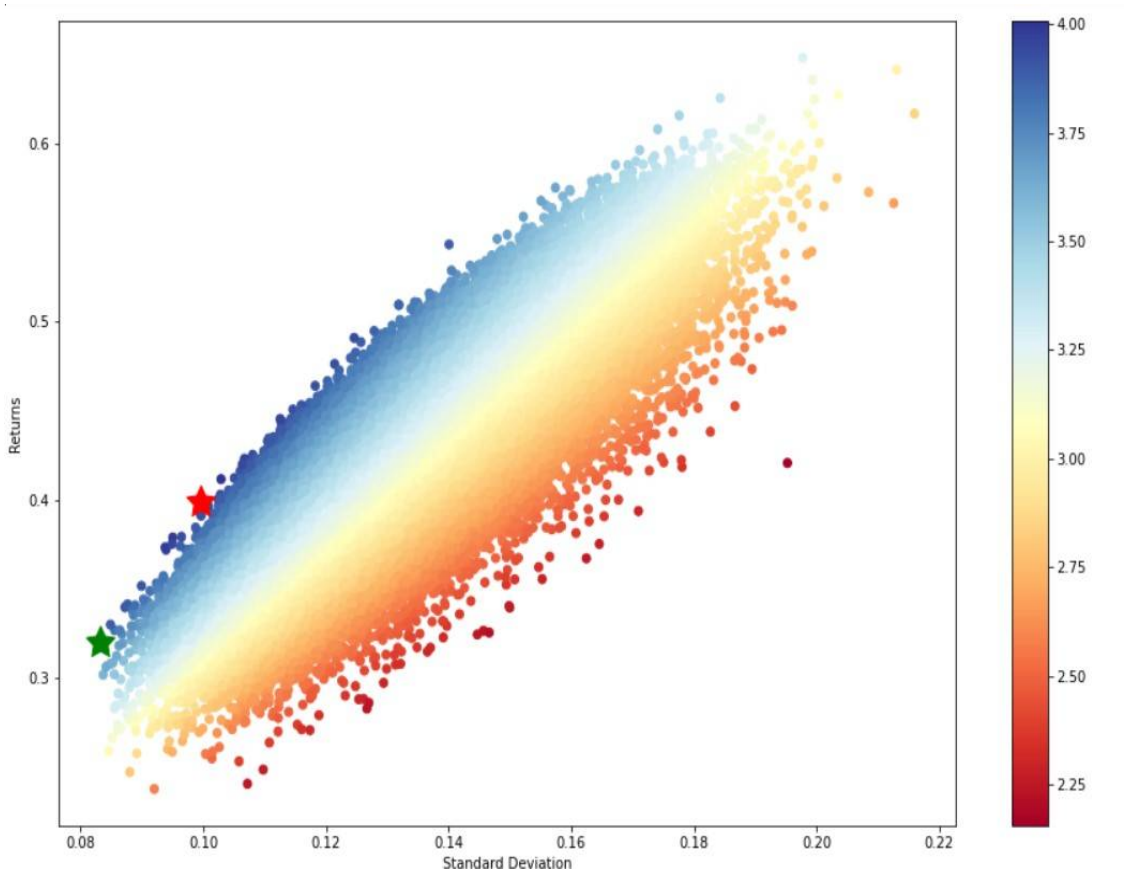


Figure 2: The Figure Represents the Efficient Frontier in the Plane Annual Return Vs Annual Volatility. The Red Star Represents the Maximum Sharp Ratio and the Green One Represents the Global Minimum Variance Portfolio During the First Period

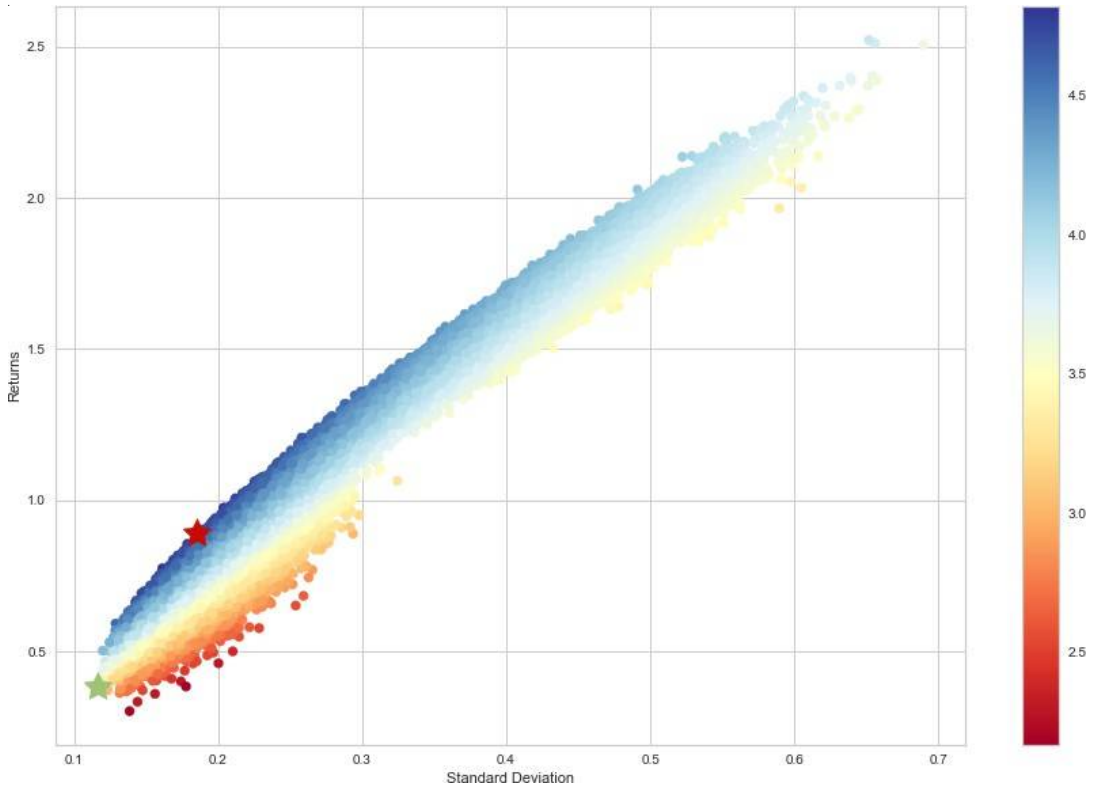


Figure 3: The Figure Represents the Efficient Frontier in the Plane Annual Return Vs Annual Volatility. The Red Star Represents the Maximum Sharp Ratio and the Green One Represents the Global Minimum Variance Portfolio During the Second Period

	sharpe	ret	stdev	Arab Tunisian Bank	Sotipapier	Soc. Tun. de Reassurance	El Wifack Leasing	AIR LIQUIDE Tun	Amen Bank	Sotumag	...
3606	4.860487	0.901049	0.185382	0.052995	0.051478	0.04997	0.049125	0.045347	0.04448	0.04168	...

1 rows × 40 columns

	sharpe	ret	stdev	Compagnie Int. de Leasing	Bq De tunisie et des Emirats	Sotumag	Placements de Tunisie	Soc. Immob. et de part.	Sotipapier	El Wifack Leasing	...
89115	3.174895	0.37305	0.1175	0.053876	0.049064	0.048616	0.047403	0.046676	0.046182	0.045279	...

1 rows × 40 columns

Figure 4: The Maximum Sharp Ratio is Represented in the Upper Row Where We Expose the Sharp Ratio, The Return, The Standard Deviation and the Weights of the Stocks with Positive Returns. The Lower Row Shows the Minimum Variance Portfolio During the Covid Crisis Starting from March 2020

portfolio and the green star represents the Global Minimum-Variance Portfolio which equals the Minimum Variance Portfolio that minimizes the risk.

This study has enabled us to make some important observations:

- The Tunisian market is characterized by a yield varying between 31% and 90% and a Sharpe ratio varying between 3.17 and 4.86. These extreme values characterize a very risky market whose data require a fine analysis. In this respect, the difficulties encountered during the application of the machine learning models will be explained below.

- The execution of the CAPM model requires a high computation time despite the fact that the number of companies is limited to 85. If we add the fact that this execution requires the calculation of a covariance matrix and its inversion during a quadratic optimization problem, this calculation becomes very costly in the case of the NYSE stock exchange with an SP 500 index containing more than 500 companies or in the case of decentralized finance for cryptocurrencies.

3.1.2. Empirical Diversification

Market finance is not a purely mathematical discipline. Indeed, in the literature there are several so-called empirical or naive diversification techniques:

- Equally Weighted (EW) portfolio.
- Cap-Weighted (CW) portfolio.
- Value-Weighted (VW) portfolio.
- Black-Litterman model.

In using these portfolios, investors do not seek to maximize performance in the most efficient way but rather to build a well-balanced portfolio where no one asset is more important than the other.

Finally, should we think about näýve diversification versus scientific diversification as a way to build portfolios? Or should we work with both approaches?

We can see that there is a wide choice of portfolios but the message has been passed that the two approaches to diversification mentioned above are not in opposition to each other but both work better when combined in the case of the theories developed after the 2008 subprime crisis. Another important message is that there is no reason why we should not look for other diversification techniques, this time using machine learning models and interpreting their results appropriately.

3.1.3. New Diversification Strategies After the 2008 Crisis

After the 2008 subprime crisis, a new concept has appeared. Indeed, classical techniques have shown their limitation in achieving a good diversification in times of crisis. For example, Finance professors have begun to talk about the use of the Constant Proportion Portfolio Insurance (CPPI) procedure allows obtaining option-like (convex) payoffs without actually using options. Liability Driven Investing (LDI): Performance-Seeking Portfolio and Liability-Hedging Portfolio (PSP/LHP). We could believe that a judiciously chosen combination of näýve and scientifically based diversification techniques would allow investors to more effectively reap its benefits across and within asset classes.

Discussion: At this point we continue the comparison between the different diversification techniques: scientific, empirical, mixed and our approach based on machine learning:

- The portfolio that maximizes the Sharpe ratio requires the calculation of the inverse of a socalled covariance matrix. In some cases, this matrix is very costly from a time complexity point of view. In a high-frequency trading context, this method becomes obsolete and inefficient. Diversification using unsupervised learning with the k-means algorithm is very fast and less expensive in terms of elementary operations. The k-means algorithm is known to have a time complexity of $O(n^2)$. Compared to the time complexity of the inversion of a matrix varying between $O(n^3)$ and $O(n^2\log(n))$ (Amund , 2003).
- Surprisingly, the literature shows that the variance minimizing portfolio is “outperformed by” the naive portfolio which consists of putting equal contributions into all companies (Victor *et al.*, 2009).
- Despite these relatively good results, the equally weighted portfolio does not allow the detection of correlated behaving assets. It is then sensitive to the systematic risks encountered during international crises. In machine learning, we speak about correlated explanatory variables. Hence the need to properly prepare the data and remove the mutually correlated variables for a better risk assessment.

3.2. Towards Optimal Diversification With the Help of Machine Learning

Before talking about diversification strategies based on machine learning, we would like to introduce you to this fast-growing field.

3.2.1. Machine Learning

The most recent technologies such as "Big Data", artificial intelligence, and "Data Science", have tried to respond to business issues through the integration of systems for the extraction, transformation, and management of large data flows, and consequently the creation of prediction models. Hence "Machine Learning", is simply a toolbox of various methods to be used in order to employ machine learning techniques contributing to model the data for the purpose of making strategic decisions. Learning models are classified into four categories: supervised, unsupervised, semi-supervised, and reinforcement.

Supervised Learning

Supervised learning consists of retrieving so-called annotated data from their outputs to train the model, i.e., associating a target label or class, and that the algorithm becomes capable, once trained, of predicting this target on new non-annotated data. This type of training is divided into two subcategories, regression, and classification.

Unsupervised Learning

In contrast to supervised learning, the unsupervised context is where the algorithm has to operate from examples with only input data and no corresponding output variables. The objective of unsupervised learning is to model the structure in the data in order to learn more about it. It is called unsupervised learning because, unlike supervised learning, there is no correct answer and no teacher. The algorithms are left to their own mechanisms to discover and present the interesting structure in the data. Below is a list of some unsupervised machine learning algorithms: K-means clustering, Density-based scan, Principal Component Analysis (PCA), Affinity propagation, Spectral clustering, Ward hierarchical clustering, Agglomerative clustering, BIRCH, Gaussian mixtures, *t*-distributed Stochastic Neighbor Embedding (*t*-SNE), Uniform Manifold Approximation and Projection (UMAP) for dimension reduction and dimensionality reduction.

Business Understanding

In the following, the objectives of this study are announced.

Determination of Business Objectives:

- Segmentation of different companies and finding the information hidden in the data.
- Create a decision support system in the Tunis Stock Exchange based on several variables deduced from finance.

Determination of Machine Learning Objectives:

- Application of different dimensionality reduction and segmentation algorithms on historical data.
- Evaluation of segmentation algorithms with metrics defined in machine learning and using an assessment based on information shared on sites specializing in finance.

3.2.2. Diversification Using Machine Learning

Unsupervised learning has several characteristics:

- Absence of a target variable.
- Provides immediate results for interpretations and later usage.
- Seeks to discover facts and knowledge concealed in large data volumes.
- This information and knowledge is already there. Above all, we are not trying to predict the future.
- This learning process allows vast data to be minimized, summarized and synthesized.

In order to prevent the curse towards high dimensions, unsupervised learning involves the reduction of dimensionality. Furthermore, it performs association rules, recommendation systems, and clustering. It is the latter technique that our article focuses on. Indeed, a segmentation problem consists of searching for the similarity between statistical observations which is in our case the Tunisian stock market companies. These observations are defined by explanatory variables (features) in an abstract space of equal size to the number of these features. Thus, the aim of segmentation is to separate statistical observations into groups called clusters by maximizing intra-group similarity and minimizing inter-group similarity (Ward's criterion). In this respect, one can easily encounter the problem of segmentation in market finance throughout the investment cycle. Traditionally, one seeks to group together companies with high profit and low risk in the Risk/Return plan. Until now, the majority of asset portfolio managers have used the traditional techniques cited above in section (3.1).

3.2.3. The Hierarchical Risk Parity (HRP)

This diversification technique was first introduced in 2015 by Professor Marcos Lopez de Prado. He believes that machine learning models and algorithms can find patterns in financial data that only business experts are capable of finding. This algorithm (L´opez de Prado, Marcos, 2019) solves 3 problems encountered in portfolios constructed based on Markowitz theory, which are instability, concentration, and underperformance:

- Concentration: These techniques can result in portfolios that are highly concentrated on a few companies in the market.
- Instability: Instability is sometimes due to poor conditioning of the covariance matrix. Indeed, conditioning is defined by the ratio of the largest eigenvalue divided by the smallest eigenvalue. If the number of conditioning is large, it means that the problem is ill-posed. This character comes from the quality of the actual data. In machine learning and during the data preparation stage, attempts are made to improve the quality of the data with linear imputation and rescaling techniques that allow the parameter space to be well prepared to facilitate the extraction of information using unsupervised learning algorithms.
- underperformance: Instability is sometimes due to underperformance. In Victor *et al*. (2009) the authors show that the out-of-sample performance for naive portfolios obtained with 1/N outperforms portfolios based on optimization techniques. Indeed, the inversion of the covariance matrix can lead to large errors that encompass the benefits of diversification.

The HRP algorithm applies the well-known Hierarchical Cluster Analysis (HCA) algorithm to build diversified portfolios using the information that resides in the covariance matrix. However, unlike Markwitz's theory based on constrained nonlinear quadratic optimization, HRP does not require the invertibility of the covariance matrix. This is a major improvement in terms of time complexity during heavy numerical computations.

4. Data Collection

After having multiple discussions with various financial experts in Tunisia, we start getting the data of the Tunisian stock market. This market contains more than 80 companies that vary on 12 multiple sectors (Consumer goods, Automobile Part, Industrials, Building Construction Basic Materials, Financials, Banks, Financial Service, Consumer Services, Tunis Food and Be, Tunis Insurance and Tunis Distribution) show in the figure below.

Nom	Dernier	Var.	Var.%
Consumer Goods	8.629,78	-21,02	-0,24%
Automobile & Part.	1.352,38	-17,11	-1,25%
Industrials	1.499,30	+9,61	+0,65%
Building Construc.	739,40	+10,82	+1,49%
Basic Materials	2.959,32	-0,58	-0,02%
Financials	4.106,89	-26,36	-0,64%
Banks	3.759,05	-25,99	-0,69%
Financial Service.	5.151,97	-34,40	-0,66%
Consumer Services.	2.674,75	-14,14	-0,53%
Tunis Food and Be.	11.050,28	-32,00	-0,29%
Tunis Insurance	12.464,79	+27,71	+0,22%
Tunis Distributio.	4.207,35	-23,77	-0,56%

Figure 5: Different Sectors of the Tunisian Stock Market

We started collecting the data by using the information provided on the Investing.com site. This platform provides all the information needed for financial analysis at this level. Using python we started scrapping the table of components having the list of companies enlisted in the Tunisian stock market and of course getting the historical data of shares in

Top Hausses »

Nom	Dernier	Var.	Var.%
Societe Moder...	1,00	+0,05	+5,26%
Societe Essou...	2,880	+0,080	+2,86%
Societe Tunisi...	3,39	+0,09	+2,73%
Telnet Holding	8,39	+0,22	+2,69%
SOTETEL	4,98	+0,08	+1,63%

Top Baisses »

Nom	Dernier	Var.	Var.%
Universal Auto...	2,93	-0,18	-5,79%
Soc. Gen. ind. ...	1,50	-0,07	-4,46%
Ateliers Mecan...	0,93	-0,04	-4,12%
Electrostar	1,300	-0,050	-3,70%
Tunisie Leasing	9,70	-0,29	-2,90%

Figure 6: Data Collected from investing.com

order to create our features later. The scrapping is done in real-time in order to get the latest information on stocks. The data we worked on starts from Mars 2019.

5. Diversifiable Risk Indicators and Data Preparation

We collected two types of data: structured and unstructured. The structured data corresponds to the quotations of the Tunisian stock market. This digital data contains the following indicators

- **Stock market index:** A stock market index is an important indicator that determines the performance of a stock market. Composed of a group of shares, it represents a market, a sector of activity. It shows the trend of the economy and of the largest companies in the index. In Tunisia, there are two stock market indices: Tunindex which represents all listed Tunisian companies, and Tunindex20 which represents the twenty most liquid listed companies with the largest market capitalizations.
- **Asset price:** The price of an asset at any point in time (daily basis) is the price at which the greatest number of securities can be traded at that point in time.
- **Opening price:** The opening price of an asset is the first price displayed at the beginning of a trading day. Generally calculated using the fixing technique, it is calculated by comparing buy and sell orders to extract an appropriate equilibrium price in the event that the maximum number of assets are traded.
- **Closing price:** The closing price of an asset is the last price fixed for that asset on the day of listing. It is considered to be the reference price for tax purposes and is one of the key pieces of information that enables investors to value their investments on the stock exchange.
- **Volume:** The volume represents the number of securities traded over a certain period of time (one trading day in our case).

The figure below shows the correlation between the 4 prices of a stock.

The figure shows a strong linear correlation with a Pearson's linear correlation coefficient of 0.99 so we can deduce that the 4 prices contain almost the same information. We, therefore, choose to work only with the close price. To include the information from liquidity in the Tunisian market we used the Volume-weighted average price **vwap** which merges the information contained in the volume and the closing price. **Vwap** is the ratio of the value traded to the total volume exchanged over a particular time horizon. In our case, we chose five days which permitted us to measure the average price of a traded share.

$$P_{VWAP} = \frac{\sum_j P_j.V_j}{\sum_j V_j} \quad ...(5)$$

where

P_{VWAP} is Volume Weighted Average Price.

P_j is price of trade j.

V_j is quantity of trade j.

Yield: Using the closing price, we calculated the linear yield time series defined by the following formula:

$$R_i = \frac{(P_i - P_{i-1})}{P_{i-1}} \quad ...(6)$$

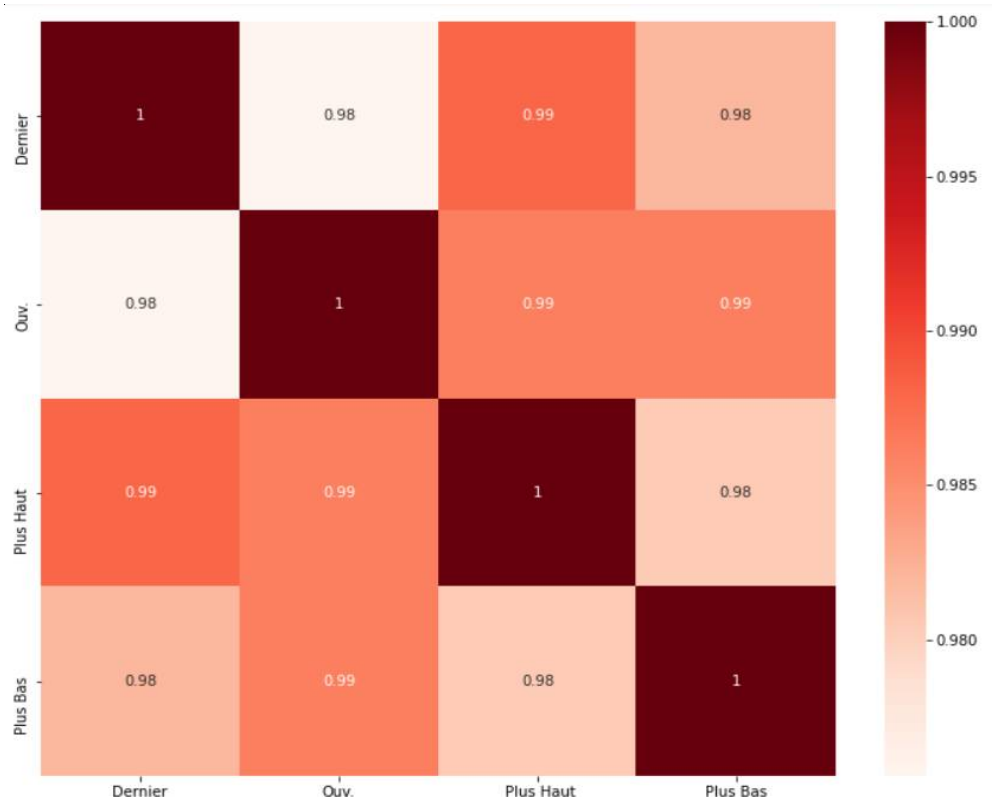


Figure 7: Correlation Between the 4 Prices of a Stock

In practice, the logarithmic yield is used in addition to the other diversification variables scaled differently.

A positive return indicates a profit while a negative return indicates a loss. From this important financial time series, the other explanatory variables (or features) are constructed:

- **Annual Yield**: The mean value of the yearly return of a certain stock.
- **Annual and Daily Volatility**: For a given security or market index, volatility is a statistical indicator of the dispersion of returns. The higher the variance, the riskier the profit, in most instances. Volatility is sometimes calculated as either the standard deviation from the market index or the difference between returns.
- **Skewness**: Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. It is assumed to be bent whether the curve is moved to the left or to the right. As a representation of the degree to which a given distribution differs from a normal distribution, skewness can be quantified.
- **Kurtosis**: Kurtosis, like skewness, is a statistical measure used to characterize distribution. Although skewness distinguishes extreme values in one tail from the other tail, in either tail, kurtosis measures extreme values. Large kurtosis distributions exhibit tail data exceeding the normal distribution tails. Low kurtosis distributions exhibit tail data that is usually less extreme than the normal distribution tails.
- **VaR**: Measures and quantifies the level of financial risk within a company, portfolio, or position over a given period of time. This metric is most often used by investment and commercial banks to determine the relationship between the extent and occurrence of potential losses in their institutional portfolios.
- **CVaR**: Also known as the estimated shortfall, Conditional Value at Risk (CVaR) is a risk assessment metric that quantifies the amount of tail risk an investment portfolio has.

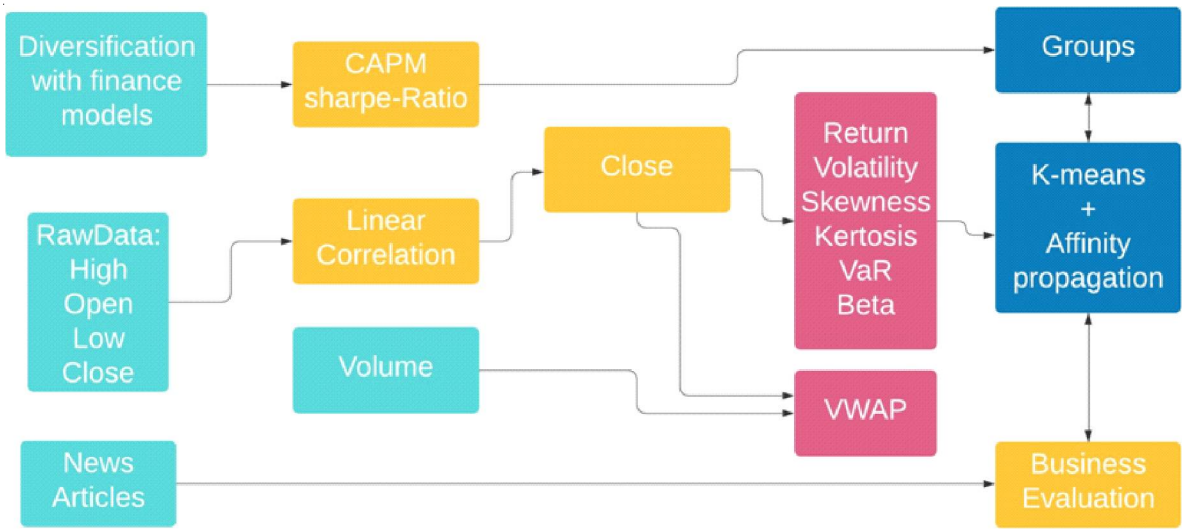


Figure 8: Analysis Architecture

- **Beta**: This coefficient measures the volatility of a security in relation to market fluctuations and therefore measures non-diversifiable risk, which makes it complementary to diversification methods aimed at minimizing diversifiable risk.

In this part, we have an extraction of features approach common in machine learning i.e. using the closing price and the vwap to calculate the time series yield, and from this series, we have calculated all the other explanatory variables. Thus each variable was calculated with the close price and the vwap. Then, a features selection step is necessary during which we started by detecting the most correlated variables two by two in order to reduce the discrepancies. We have applied a PCA and the *t*-SNE algorithm.

During the correlation study, both Pearson and Spearman correlation matrices were calculated. The linear correlation matrix presented on the left side of figure 9 shows a strong linear correlation between annual return and annual volatility. Surprised by this result, we checked the scatterplot in the yield-volatility plane and found that an outlier (the STEQ company: this result will be confirmed using the outlier detection algorithms) is just localized at the level of the first bisector which explains this artifact. By removing this outlier, we notice that this strong correlation has disappeared as shown in figure 9 on the right. Figure 10 shows a non-linear Pearson correlation matrix. At the end of this study, we notice

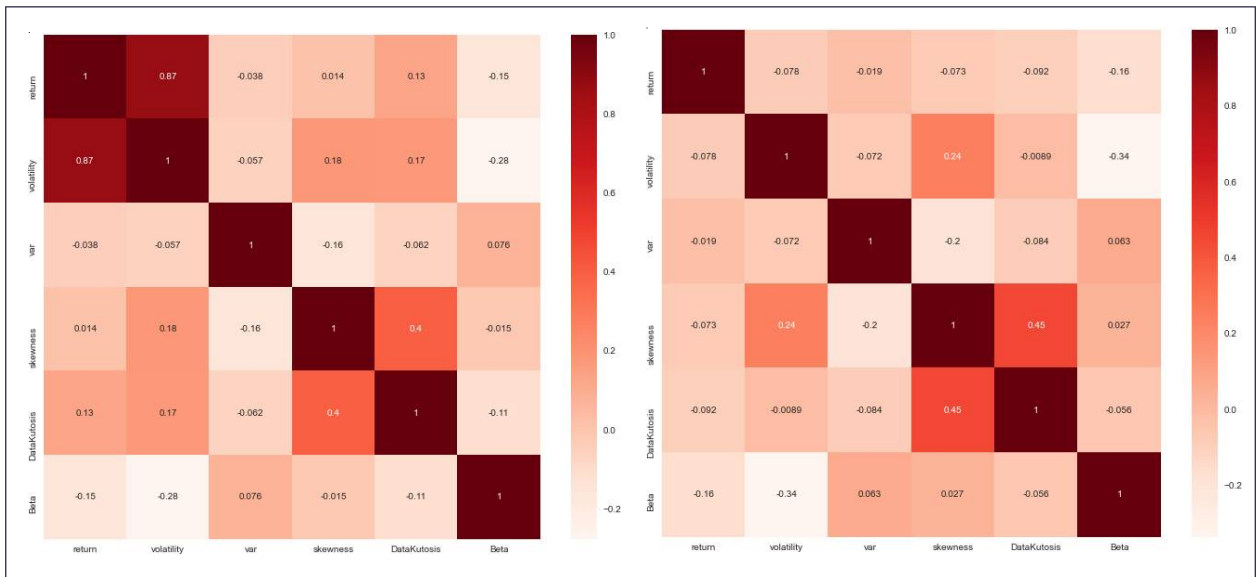


Figure 9: Correlation Matrixes Using Pearson

the inexistence of strong mutual correlations which justify the suppression of some features. We also note that the features calculated from the closing price are almost correlated to the features calculated using the vwap. In this respect, we decided to work only with features based on the closing price.

	X_1	X_2	...	X_p
c_1	x_{11}	x_{21}	...	x_{p1}
c_2	x_{21}	x_{22}	...	x_{21}
...	...	...	...	...
c_n	x_{1n}	x_{2n}	...	x_{pn}

6. Application of Unsupervised Learning

Segmentation models focus on data tables with n companies $n = 85$ companies with $c_1, c_2, \cdots, c_n$ in rows and $m \in [2, 14]$ numerical features $X_1, X_2, \cdots, X_p$ in columns.

Objectif: These models try to group the$(c_i)_{1d\ id\ n}$ companies that are most similar by constructing different groups based on similarity. This notion is deducted using the calculation of various distances such as Euclidean, Manhatan and Minkowski distances.

6.1. K-means Algorithm

6.1.1. Mathematical Concepts

This is the best-known algorithm in unsupervised learning and is used in segmentation. The k-means algorithm is used to partition a given set of observations into a predefined amount of *k* clusters. The algorithm starts with a random set of *k* center-points (μ). During each update step, all observations *c* are assigned to their nearest center-point (see equation 7). In the standard algorithm, only one assignment to one center is possible. If multiple centers have the same distance to the observation, a random one would be chosen.

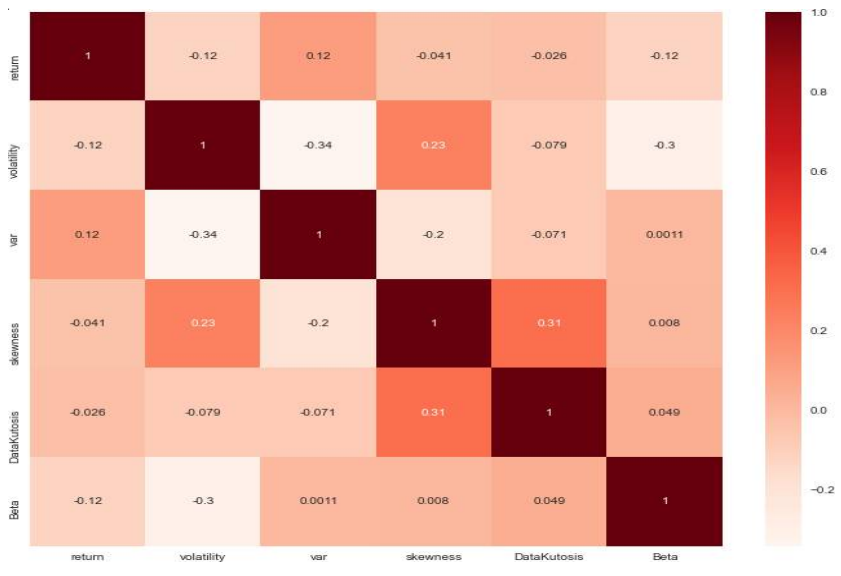


Figure 10: Correlation Matrix Using Spearman Method

$$S_i^{(t)} = \left\{ c_p : \left\| c_p = \mu_t^{(t)} \right\|^2 \le \left\| c_p = \mu_t^{(t)} \right\|^2 \forall j, 1 \le j \le k \right\} \quad ...(7)$$

Afterward, the center-points are repositioned by calculating the mean of the assigned observations to the respective center-points.

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{c_j \in S_i^{(t)}} c_j \quad ...(8)$$

The update process reoccurs until all observations remain at the assigned center-points and therefore the center-points would not be updated anymore.

This means that the k-means algorithm tries to optimize the objective function 9. As there is only a finite number of possible assignments for the number of centroids and observations available and each iteration has to result in a better solution, the algorithm always ends in a local minimum.

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|c_i - \mu_k\|^2 \quad ...(9)$$

$$\text{with } r_{ik} = \begin{cases} 1 & c_i \in S_k \\ 0 & \text{otherwise} \end{cases}$$

The main problem of k-means is its dependency on the initially chosen centroids. The centroids could end up splitting common data points whilst other, separated points get grouped together if some of the centroids are more attracted by outliers. The most common approach is to perform multiple clusterings with different start positions. Afterward, the clustering, which occurred most is considered correct. This problem is solved with a newer approach is the so-called k-means++ (David and Sergei, 2007). This extension to the k-means algorithm tries to distribute the initial centroids over the given data to minimize the probability of bad outcomes. The initial points are set by the following steps:

1. Take uniformly a random data point from the data X and mark it as centroid μ_1
2. Choose another centroid μ_k with the probability $\frac{D(c)^2}{\sum_{c \in X} D(c)^2}$ where $D(c)$ denotes the shortest distance from the data point c to its closest, already chosen centroid.
3. Repeat 2. until all k initial centroids are chosen.

Afterward, the standard k-means algorithm as described above is performed. It has been shown that with this initialization algorithm, k-means++ approximately can be computed in $O(\log n)$, compared to $O(n^{dk+1} \log n)$ for the standard algorithm. In the following paragraphs, the evaluation metrics used to validate the results obtained after the application of the segmentation algorithms are presented.

6.1.2. K-Means Applied on Tunisian Financial Problems

Clustering problems occur naturally in finance, at every stage of the investment process. The first objective of the paper will be to discuss the general theory of segmentation in the case of financial data. The second and main objective is to build homogeneous and above all explainable groups of Tunisian companies. The third objective is to compare these results with financial news from the “ilboursa.com” website. During this analysis, we have applied the segmentation algorithms for implementation via stable libraries that are available on GitHub. The kmeans algorithm gave the best results in terms of evaluation and comparison with the algorithms for detecting outliers. A particular interest was given to the reproducibility and interpretability of the results.

However, separating companies into groups by maximizing intra-group similiraties and minimizing inter-group similarities is contradictory by definition (Shai, 2018).

Let’s discuss the gap between theory and practice: Clustering is one of the most widely used tools for exploratory data analysis: social sciences, biology, astronomy, computer science... All of them apply clustering to gain a first understanding

of the structure of large data sets. Unfortunately, however, there is unfortunately little theoretical knowledge about clustering.

Lack of clustering theory: When choosing a course or a textbook for machine learning, we can find a huge difference between learning theory and unsupervised learning theory. For example, in learning theory, we can find sound theoretical principles such as bias/variance trade-off, Chernoff/Hoeffding limits, VC dimensions, practical advice on how to use learning algorithms?

On the other hand, we get a long list of algorithms K-means, EM, Gaussian mixture, factor analysis, PCA, ICA (Independent Component Analysis), There are no principles, no guidelines, just a long list of algorithms. To this end, we found major difficulties in obtaining interpretable results during our analysis. Indeed, we encountered discrepancy in the results of the evaluation metrics.

Thus, as a segmentation strategy, we have adopted a division according to time periods:

1. **First Segmentation Between 1/1/2019 and 1/3/2020:** This is a period that preceded the spread of viruses around the world. The goal is to see the state of the Tunisian market during a normal period with enough statistics since we will be working on 310 trading days.
2. **Second Segmentation from 1/3/2020 to 30/12/2020:** This is the period when Tunisian enterprises were hit hard by the Covid-19 crisis. The aim is to quantify the effect of this crisis on the distribution of enterprises.
3. **A Classical Analysis Using Two Traditional Variables:** Average annual return and average annual volatility. The aim is to align with the analyses carried out by financiers using the CAPM model.
4. **A Focused Analysis:** Using a greater number of decision variables (features). We used a total of 7 features. The idea is to follow the machine learning methodology, which consists of carrying out feature engineering to construct the largest number of variables, thus rationally increasing the parameter space and then letting the algorithms find the patterns.

6.2. Silhouette Scores

Silhouette scores are defined for each sample $\{S_i\}_{i=1,\dots n}$

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)} \quad \dots(10)$$

where

a_i is the mean distance between an object *n* and another object in its cluster.

b_i is the mean distance between object *n* and the other objects in the nearest cluster.

The advantages of the Silhouette index are:

The scores are bounded $S_i \in [-1, 1]$

We are able to reallocate specific objects to more suitable clusters because we have one score per sample.

The clusters with an average $S_i \approx 0$ could be merged.

6.3. Calinski-Harabasz Index

Given *n* companies, centered around μ.

the within cluster dipersion W_k is:

$$w_k = \sum_{k=0}^{K} \sum_{i \in k} (c - \mu_k)^T (c - \mu_k) \quad \dots(11)$$

where C_k is the set of objects in cluster *k*, μ_k is the center of the cluster *k*, and *c* is the coordinates of the object.

We define another variable called te between cluster B_k is :

$$w_k = \sum_{k=1}^{K} \sum_{i \in k} N_k (\mu_k - \mu)^T (\mu_k - \mu) \quad \dots(12)$$

where N_k is the number of objects in C_k.

Thus, the variance ration is defined as:

$$S = \frac{B_k}{W_k} x \frac{N-K}{K-1} \quad \text{...(13)}$$

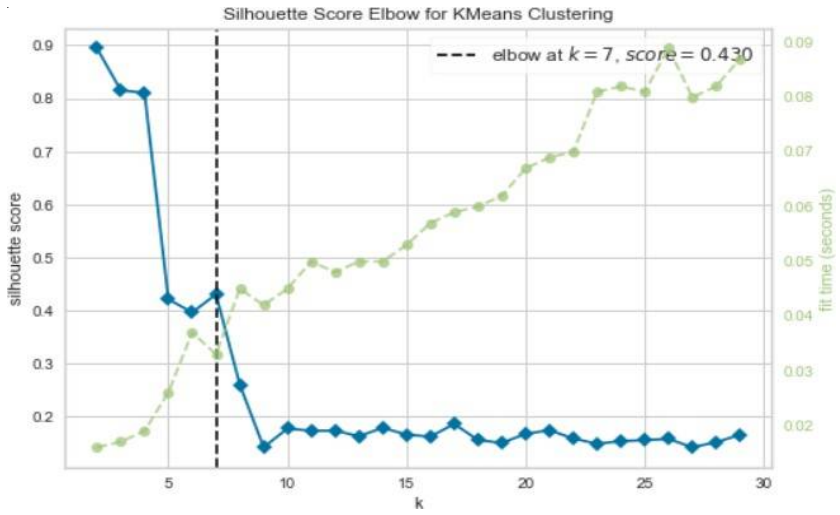


Figure 11: Distortion Silhouette for Clustering Shows the Presence of 7 Groups

6.4. Elbow Method

The Elbow method is used to determine the optimal value of *k* for clustering. The fundamental principle behind this technique is that by changing *k*, it plots the different cost values. The elbow point, which functions as an ideal value of *k*, is the point where this distortion decreases the most. For different *k* values, measure the Within-Cluster-Sum of Squared Errors (WSS) and pick the *k* for which WSS becomes the first to decrease. This is visible as an elbow in the plot of WSS-versus-*k*. Squared Errors Within-Cluster-Sum sounds a bit complicated. We're going to break it down:

- For each point, the Squared Error is the square of the point's distance from its representation.
- For all the points, the WSS score is the sum of these Squared Errors.
- Any distance measure may be used, such as the Euclidean distance or the Manhattan distance.

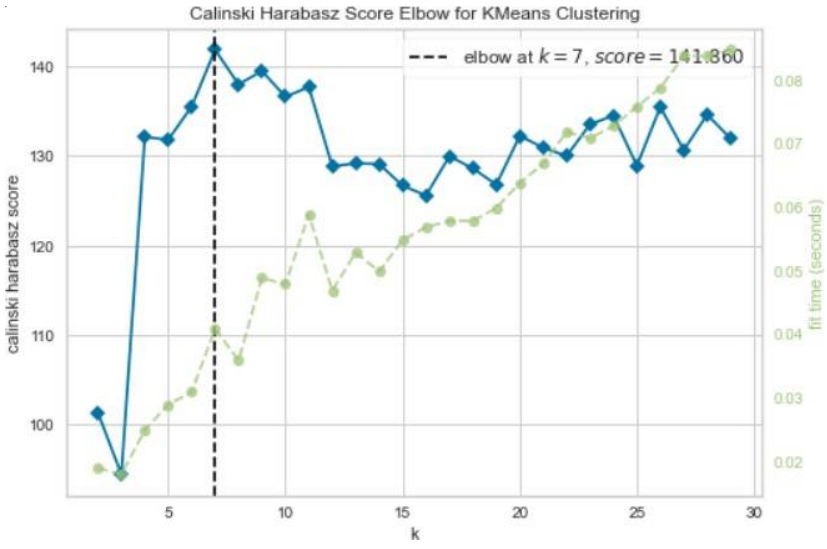


Figure 12: Distortion Calinski Elbow for Clustering Shows the Presence of 7 Groups

7. Results and Evaluation

So far, we have reviewed the different methods of diversification and construction of financial asset portfolios. In this part, we propose to apply segmentation to divide companies based on their similar characteristics in the traditional plan (annual return, annual volatility) used by finance specialists and in a space of parameters built by 6 dimensions constituted by (return, volatility, var, skewness, kurtosis, beta). These features are calculated one time with the close price and another time with vwap. It is believed that this empirical and rapid segmentation can help investors to build more efficient portfolios especially in the context of HFT (High-Frequency Trading).

7.1. Results Before Covid-19 Crisis

The application of the isolation forest algorithm gave the following results. The list of outlier companies with annual volatility and annual return variables is: MIP, STEQ, ELBEN, SITX, AST. In on the other hand the list of outliers with all variables using PCA vectors with an explained inertia of 55% contains AST, ICF, MGR, SERVI, ELBEN.

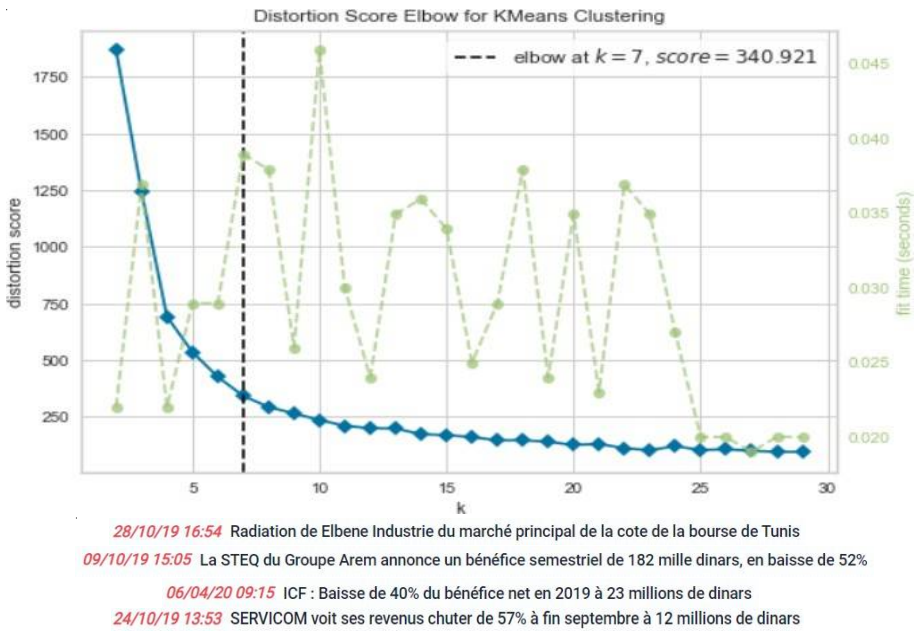
In order to confirm the results of the Isolation Forest we opted on having references from articles given in 'ilboursa.com' after developing an automatic scrapper with given keywords to detect anomalies in the downloaded results:

for the purpose of visualizing the results of the segmentation clearly after determining the outliers as mentioned before we used K-means clustering with 7 centers. In the figure 14, the outliers are clearly the ones presented at the same time as centers with the exception of SITX and AST that represent a group together but fall under the umbrella of outliers determined by the use of isolation forest.

In Figure 14 the blue group represents a group with a all the banks in our data with the exception of UBCI. This phenomena is present throughout the rest of the figures were we discuss why is that occurring and what are the references talking about this issue. After eliminating companies with highly anomalous behavior, the companies projected in the return-volatility plane are arranged in the form of a triangle. The apex on the left is formed by companies with low volatility and reasonable returns. In effect, these companies are solid. Another thing, we also find the first capitalizations like SFBT, BIAT and BT...

The top one on the right has a high yield coupled with a high volatility. Finally, the bottom top has a low return and high volatility.

It is interesting to note a large concentration of banks (private, public and foreign) in the same group. If we assume that the information retrieved from the closing price represents the state of the bank, then it is clear that a large proportion of banks behave in the same way in the return-volatility plane. This phenomenon is not new since studies published by the central bank show the interest of a future merger of its banks because on the one hand they have the same financial behavior and on the other hand in order to decrease the exposure of different risk.



13: Distortion Score Elbow for Clustering Shows the Presence of 7 Groups

In order to get more accurate results with the companies that are not considered as outliers. We opted to evaluate our segmentation with a second iteration but after removing the companies that got segmented as outliers and confirmed when getting the results of the first k-means. While concentrating on banks we found that UBCI was represented again on another cluster other than the usual group, BTEI and ATB are in a different group and the rest of the banks are clustered together.

For the next plot, we use sparse inverse covariance estimation to find which quotes are correlated conditionally on the others. Specifically, sparse inverse covariance gives us a graph, that is a list of connections. For each symbol, the symbols that it is connected to are those useful to explain its fluctuations. We use clustering to group together quotes that behave similarly. Here, amongst the various clustering techniques available in the scikit-learn, we use Affinity Propagation as it does not enforce equal-size clusters, and it can choose automatically the number of clusters from the data. Finally, for the visualization, we use Manifold learning techniques to retrieve 2D embedding.

7.2. Covid-19 Crisis Results

List of outliers with annual volatility and annual return variables: SIMP, STEQ, ALKM, SITX and TINV.

List of outliers with all variables using ACP vectors with an explained inertia of 55%: BNA, STEQ, TINV, UIB and TGH.

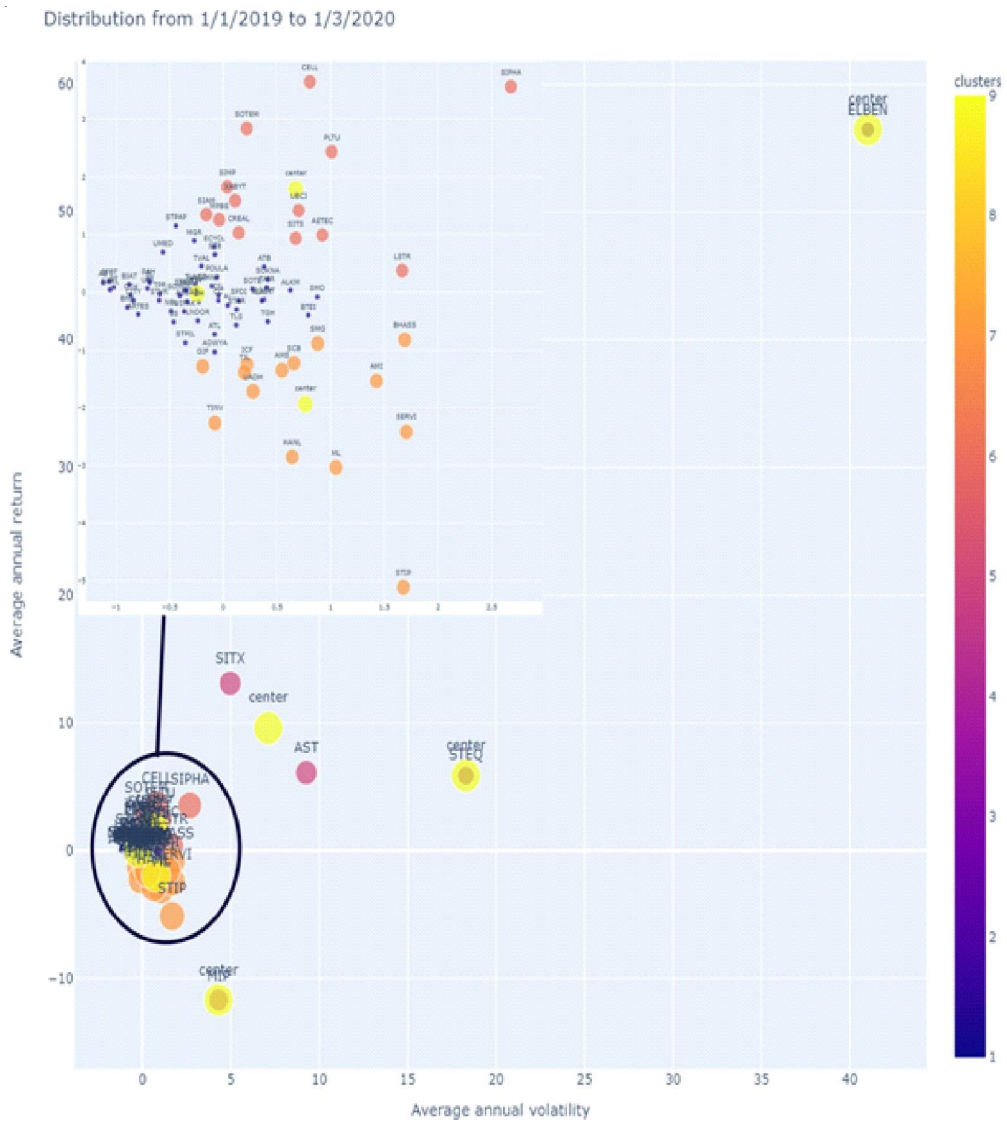


Figure 14: Results of the K-Means Clustering of with k = 7 Groups in the Plan Average Annual Return- Average Annual Volatility. Clustering Results Use Only These Two Classical Variables

Again for the year 2020, in order to confirm the results of the Isolation Forest we used to detect in details our outliers with the use of 5% contamination which represents 5% of the group of companies, we opted on having references from articles given in 'ilboursa.com' after developing an automatic scrapper with given keywords to detect anomalies in the downloaded results we got the following inputs:

7.2.1. Analysis of the Covid Crisis

In order to visualize the results of the segmentation clearly after determining the outliers as mentioned before, we used K-means clustering with 7 centers. There exist 3 groups of outliers two are coupled with the exception of STEQ.

Figure 21 shows a triangle configuration of companies. In the risk-return plane with return on the vertical axis and risk on the horizontal axis, the companies that form the left vertex of the triangle represent strong companies with low and reasonable returns. In the upper right-hand corner, we find the company STEQ. According to our analysis, this company shows an anomalous behavior compared to other companies listed on the stock exchange. This behavior has been confirmed by its delisting from the stock exchange as of January 6, 2021. It should be noted that our analyses, which have been underway since March 2020. When this company is removed, the configuration of the triangle improves. Now the same peak is formed by companies with speculative bubbles. The most striking example is the company UADH,

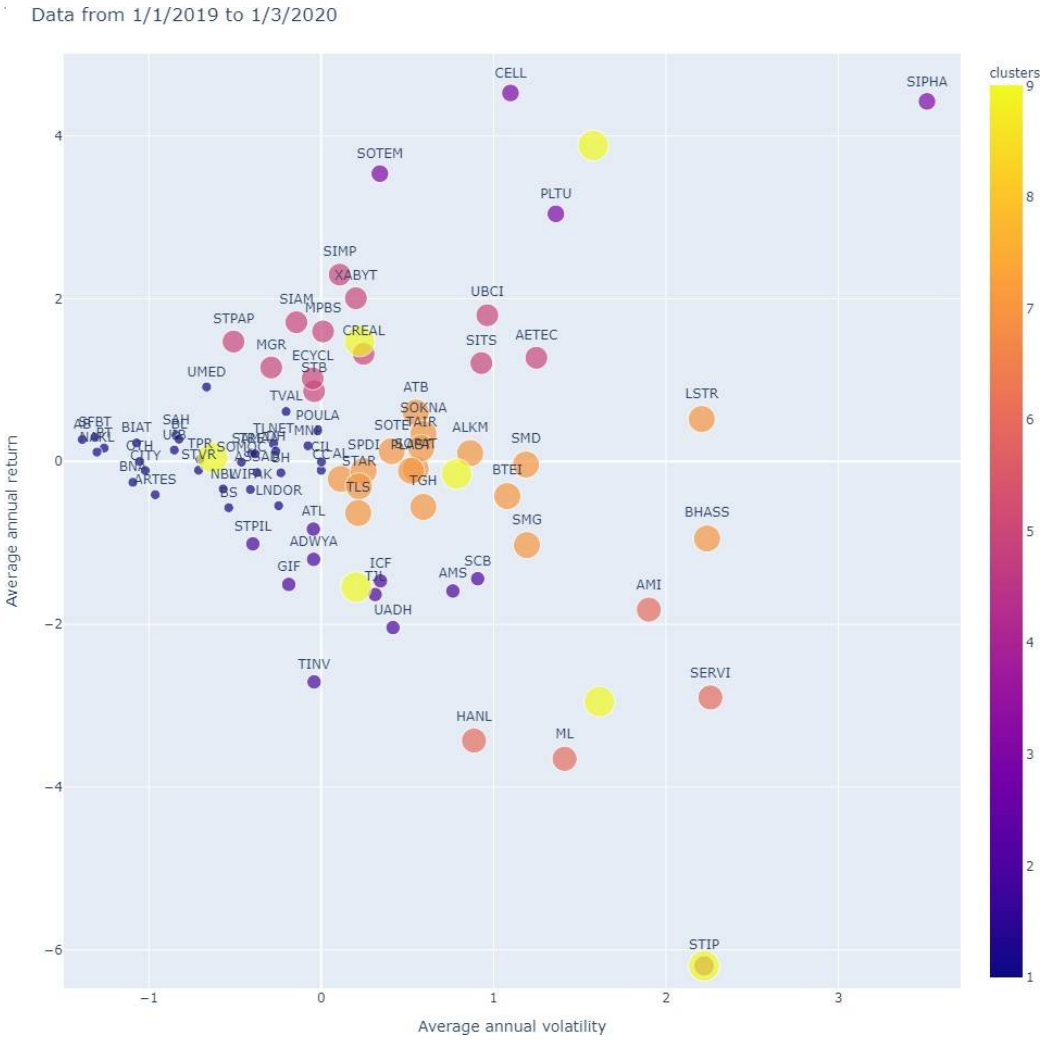


Figure 15: Second Iteration Clustering Results Using Two Variables

whose listing was suspended on March 23, 2021. In the same cluster, we find the company AMS characterized by major financial problems and consecutive business stoppages.

On March 31, 2021, the court ordered the admission of AMS under the judicial settlement procedure and the company was delisted from the stock exchange on April 5, 2021. In Figure 21 the blue group represents a group with all the banks in our data with the exception of UBCI. This phenomenon is present throughout the rest of the figures where we discuss why is that occurring and what are the references talking about this issue. The Figure 22 is obtained during the Covid-19 crisis in Tunisia so between the first of March 2020 and the end of December 2021. Clustering results with outliers with the k-means algorithms applied on the two components. These components explain 51% of the variability and are obtained with the application of PCA. The Figure 23 is obtained during the Covid-19 crisis in Tunisia so between the first of March 2020 and the end of December 2021. Clustering results without outliers with the k-means algorithms applied on the two components These components explain 51% of the variability and are obtained with the application of PCA. For the next plot, we use sparse inverse covariance estimation to find which quotes are correlated conditionally on the others. When discovering the clusters it should be mentioned that the majority of the banks are represented in group 1 mainly and the pharmaceutical companies are represented in the fourth group.

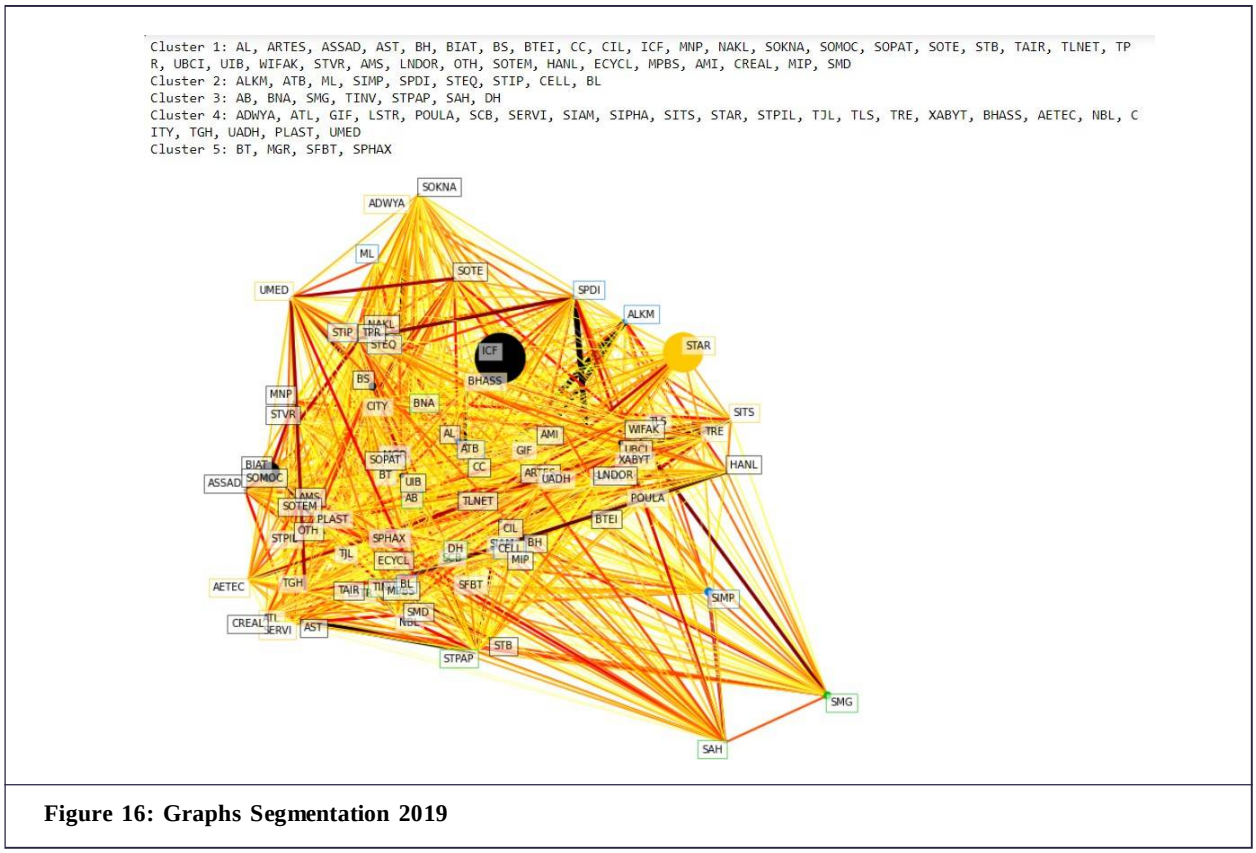


Figure 16: Graphs Segmentation 2019

8. Architecture

To cut along story short, as follows is the work structure. we start by the raw data which is the data scrapped of the Tunisian financial market. Starting with the first column of input which contains the four major inputs that are the Share high, Open, Low and close of every day in a specific interval. Added to that the news articles used for the business evaluation as a final step. For the data we got we apply a linear correlation to decide on working on the close column because of existance of high correlation between the four columns. From the close column we applied multiple operations to obtain the return, volatility, skewness, kertosis, value at risk and the Beta. These values are applied into unsupervised machine learning algorithm which is in our case Kmeans with the affinity propagation to deduce the clusters. In order to validate the accuracy of these groups we opted on working on a diversification using traditional financial models which is the case of sharp ratio. Finally to evaluate the results we used the news articles from the Net applied on key structures to deduce if our work applies on the experts financial results of each specific group.

05/01/21 14:35 Radiation de la Cote de la Bourse de la société STEQ du Groupe AREM à partir de demain
20/11/20 08:51 Le Groupe TAWASOL annonce des revenus en chute de 45% à fin septembre
21/10/20 13:06 TUNINVEST annonce des revenus des participations en hausse de 81% au troisième trimestre

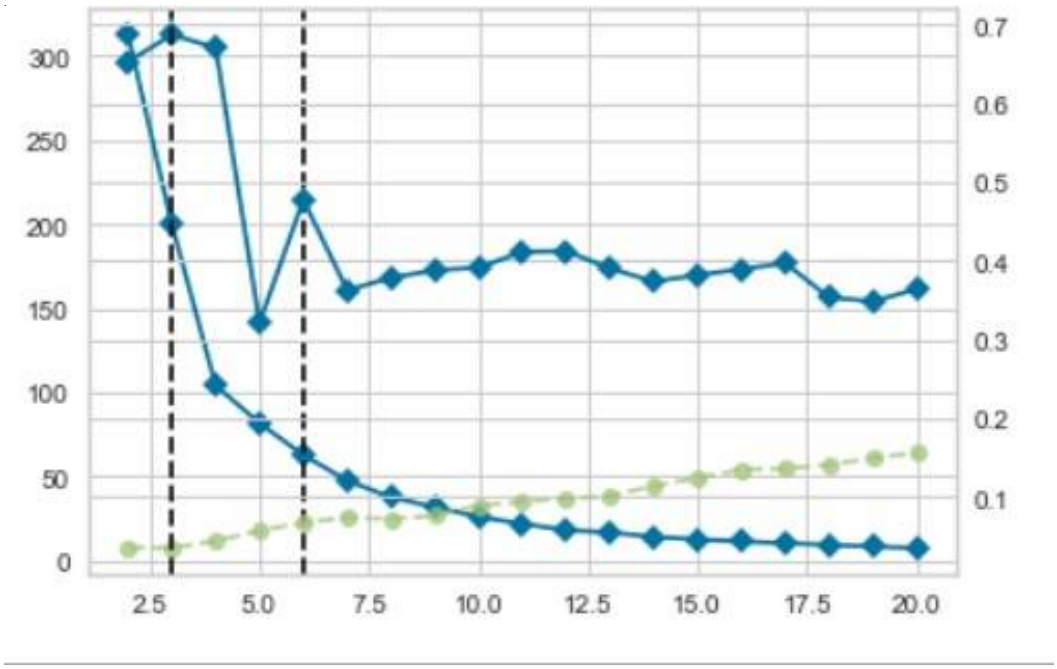


Figure 17: Distortion and Silhouette Elbows for Clustering

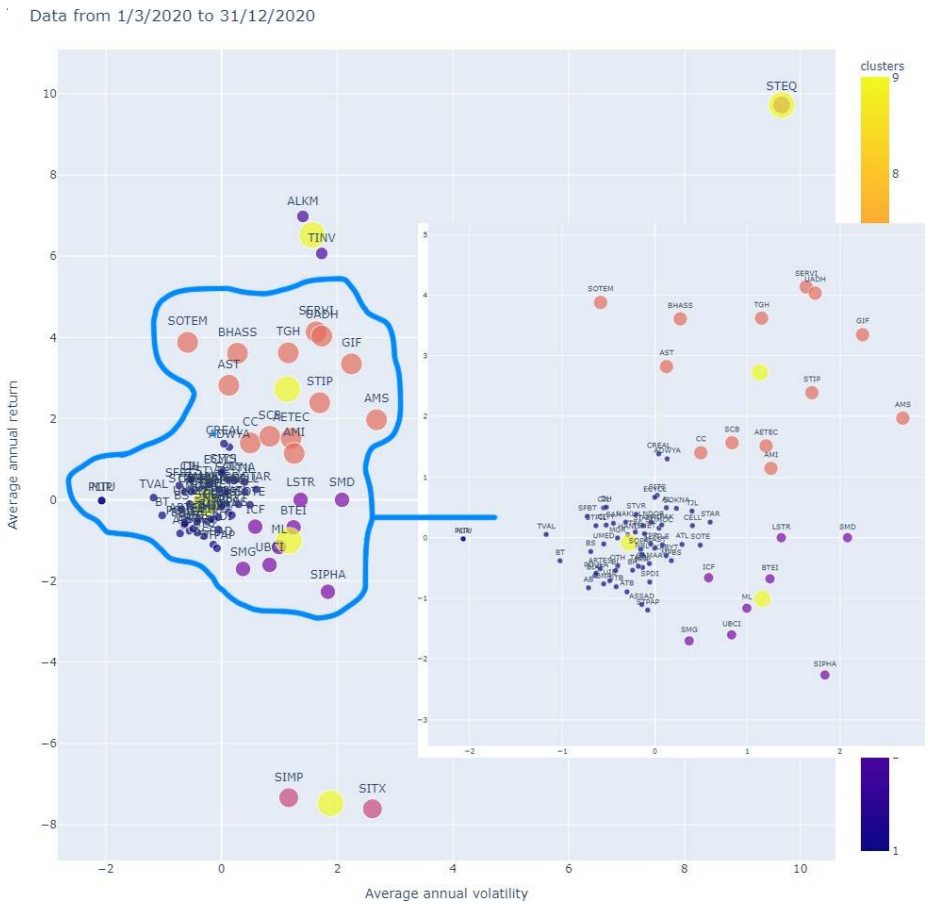


Figure 18: Clustering Results Using Two Variables

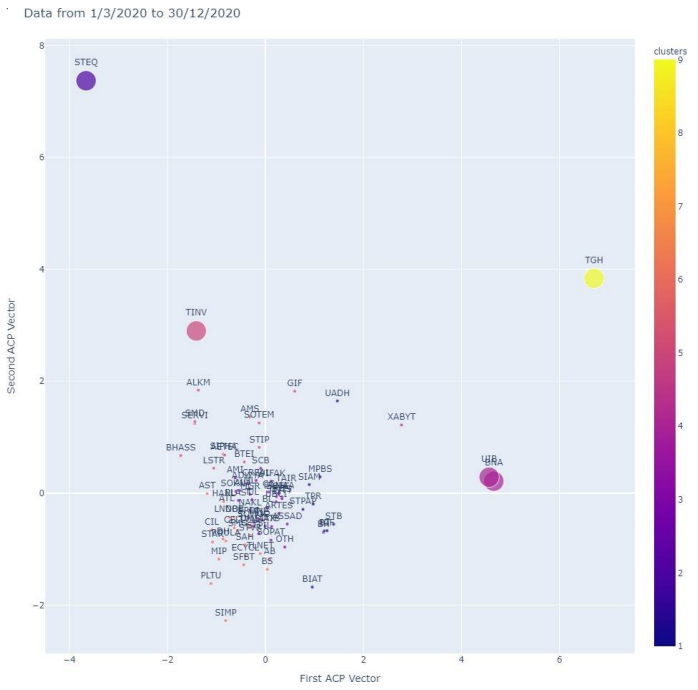


Figure 19: The Figure is Obtained During the Covid-19 Crisis in Tunisia So Between the First of March 2020 And the End of December 2021. Clustering Results with Outliers with the K-means Algorithms Applied on the Two Components These Components Explain 51% of the Variability and are Obtained with the Application of a Principal Component Analysis

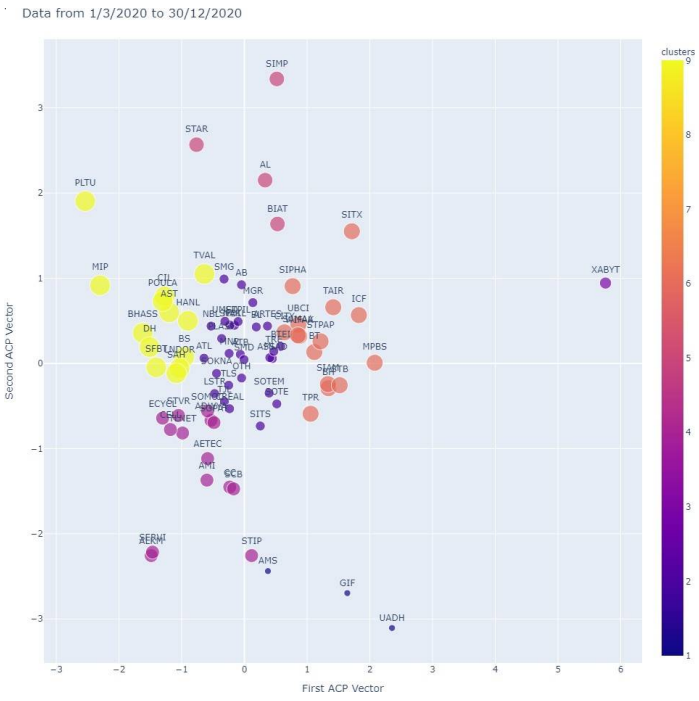


Figure 20: The Figure is Obtained During the Covid-19 Crisis in Tunisia So Between the First of March 2020 and the End of December 2021. Clustering Results Without Outliers with the K-Means Algorithms Applied on the Two Components These Components Explain 51% of the Variability and are Obtained with the Application of a Principal Component Analysis

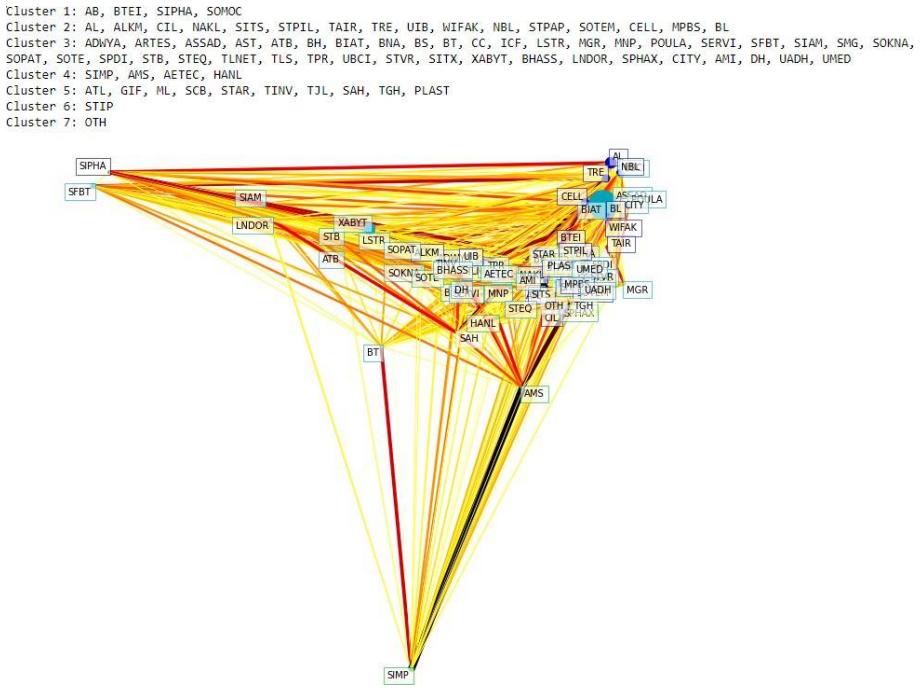


Figure 21: Graphs segmentation 2020

9. Conclusion

Finance is concerned with the decision to spend available capital in an uncertain manner over a set time span in order to make a profit. As a result, such a decision could put the allocated funds at risk. As a result, there is a need for diversifiable risk diversification since there is another incompressible factor that contributes to systemic risk (non-diversifiable). Historically, the CAPM (Capital Asset Pricing Model) theory has provided a sound theoretical foundation for achieving such diversification and obtaining asset portfolios that minimize risk while maximizing return through quadratic optimization. Unfortunately, when it comes to risk management, a portfolio that maximizes the Sharpe ratio is not always the best option, especially during extremely rare 4 sigma events like financial or health crises. This was seen during the subprime mortgage crisis of 2008 and the current Covid-19 crisis. Indeed, in a parametered space, the problem of diversifying a portfolio of financial assets can be reformulated as a segmentation problem in the same space. With its unsupervised learning methods, Artificial Intelligence provides algorithms for detecting weak signals in large amounts of data. After working on the Tunisian stock market we were able to detect groups that are valuable to invest in and on the other hand very risky groups that their stock value is unpredictable. We first used isolation forest in order t obtain the companies that are really far from the average fluctuations and to confirm the first iteration we used the clustering techniques explained previously and the results were confirmed as the companies in question were either in the same group alone or each one of them was provided a center of a cluster and occupied the group by itself. All the results we got were confirmed by news articles specializing in the Tunisian stock market.

References

Amund, Tveit. (2003). On the Complexity of Matrix Inversion. Department of Computer and Information Science, Norwegian University of Science and Technology,N-7491 Trondheim, Norway.

Baker, S.R., Bloom, N., Davis, S.J., Kost, K.J., Sammon, M.C. and Viratyosin, T. (2020a). The Unprecedented Stock Market Reaction to Covid-19. *The Review of Asset Pricing Studies*, Forthcoming.

Black, Fischer. and Scholes, Myron (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3), 637-654.

Cont, Rama and Tankov, Peter. (2007). Constant Proportion Portfolio Insurance in Presence of Jumps in Asset Prices. February. Columbia University Center for Financial Engineering, Financial Engineering Report No. 2007-10.

Carhart, M.M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52(1) 57-82. doi:10.1111/j.1540-6261.1997.tb03808.x. JSTOR 2329556

DeMiguel, V., Garlappi, L. and Uppal, R. (2007). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy. *The Review of Financial Studies*, 22(5), 25.

David, Arthur. and Sergei, Vassilvitskii. (2007). K-Means++: The Advantages of Careful Seeding. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1027-1035.

Fama, E.F. and French, K.R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), 3-56.

Fama, E.F. and French, K.R. (1996). Multifactor Explanations of Asset Pricing Anomalies. *Journal of Finance*, 55-84.

Harry, Markowitz. (1952). Portfolio Selection. *The Journal of Finance*, 7(1). March.

Imen, Boukhicha. (2020). Bank Competition and Financial Stability: A Case Study of Tunisian Banking Industry.

Jamila, Nachnouchi. (2020). Banques Tunisiennes et risque syst´emique: Approche par la CoVaR.

John, B. Rollins. (2015). Foundational Methodology for Data Science. IBM Analytics White Paper, June.

Kaizoji, Taisei and Miyano, Michiko. (2018). Stock Market Crash of 2008: An Empirical Study of the Deviation of Share Prices From Company Fundamentals. *Applied Economics Letters*, 26(5), 362-369.

López de Prado, Marcos (2019). Ten Applications of Financial Machine Learning. https://ssrn.com/abstract=3365271, September 22.

Leung, E., Lohre, H., Mischlich, D., Sheah, Y. and Stroh, M. (2021). The Promises and Pitfalls of Machine Learning for Predicting Stock Returns. *Journal of Financial Data Science*.

Maurizio, Dallocchio., Yann, Le Fur., Pascal, Quiry., Antonio, Salvi. and Pierre, Vernimmen. (2018). *Corporate Finance: Theory and Practice*, p. 20.

Perold, Andr´e F. (2004). The Capital Asset Pricing Model. *The Journal of Economic Perspectives*, 18(3), 3-24. JSTOR, available at www.jstor.org/stable/3216804. Accessed on December 28, 2020.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *J Data Warehousing*, 5(4), 13-22.

Shai, Ben-David. (2018). Clustering - What Both Theoreticians and Practitioners are Doing Wrong. available at arXiv:1805.08838

Victor DeMiguel, Lorenzo Garlappi, Raman Uppal (2009). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?. *The Review of Financial Studies*, 22(5), May.

Cite this article as: Ahmed Rebai, Louay Boukhris, Lotfi Ncib and Mohamed Anis Ben Lasmer (2021). Unsupervised Learning Diversification Applied on the Tunisian Stock Market Before and During the Covid-19 Crisis. *International Journal of Management Research and Economics.* 1(4), 24-47. doi: 10.51483/IJMRE.1.4.2021.24-47.