Faculty of Economics and Business Administration.

MACHINE LEARNING APPLIED TO ESG FACTOR INVESTING.

Author: María Loreto Delgado Fernández-Valdés.

Director: María Coronado Vaca.

MADRID | June 2023

ABSTRACT

The current financial world is experiencing a transformation and increasingly evolving towards environmental, social and governance (ESG) criteria as the future of investment strategies. This situation is a result of investors growing consciousness of environmental and social challenges, and the consequent integration of ESG aspects in their financial decisions.

This article is focused on exploring the impact of ESG criteria on factor investing. Machine learning is additionally introduced as a tool for managing large volumes of data and, thus, improve factor selection, risk management and decision making in factor investment strategies.

The results obtained in this study show a positive impact of ESG integration in factor investing. These findings not only indicate that ESG criteria are valuable from an ethical and sustainability perspective, but also demonstrate that they can yield higher financial returns. Evidence on this fact is still evolving and, as more evidence accumulates on the positive impact of ESG criteria on investment strategies, investors will probably adapt their financial decisions towards these practices.

Key words: Factor Investing, ESG, Returns, Performance, Machine learning.

INDEX

INDEX OF FIGURES

1. INTRODUCTION

1.1. Objectives

This article seeks to study the impact of adding ESG criteria as an additional factor in a factor investing strategy. To do so, the following objectives are stablished:

- Analyze the origin of Factor Investing to better understand what factor investing is.

- Additionally, it is defined and explained what factor investing is, as well as the factors that have been mostly used in previous studies.

- Once understood the concept of factor investing, a new factor will be studied, the ESG factor as a possible engine of better results.

- Analyze Machine Learning techniques that are popularly applied to Factor Investing strategies to optimize these latter ones.

- Perform a practical example to test the conclusions drawn.

1.2. Justification

Nowadays the society is increasingly becoming aware of the importance of the planet care. For example, in a survey run by the United Nations they found that the majority of the worldwide population calls for action to fight against climate change (United Nations Development Program, 2021). Furthermore, Environmental, Social and Governmental (ESG) criteria already has an impact on many aspects, including finance.

In the last few years, the investment world has been evolving towards a more sustainable approach and new investment alternatives have been developed to adapt to these trends. For instance, Carmignac, an asset management company, states in their website that their sustainable funds represent 90% of their total assets under management. Furthermore, they have developed an ESG Calculator for investors to view the non-financial results, i.e., the impact on social and environmental matters associated with their investments in Carmignac's equity funds (Carmignac, n.d.). As well as Carmignac, many other financial players have developed sustainable investment alternatives.

On the other hand, since the study of Fama & French in 1992 (Fama & French, 1992), Factor Investing has been evolving and many studies on this topic have been performed.

For these two reasons, I considered a relevant field of study the impact of ESG criteria on Factor Investing. Furthermore, as a way to optimize the management of the vast amount of data currently available Machine Learning techniques seem very useful.

1.3. Structure

The article is structured in four sections:

After this introduction to the content of the article, the main topic, Factor Investing, is studied in the second section. Therefore, the origin and main concepts of factor investing will then be explained. To complete the knowledge and understanding of factor investing, the most popular factors as well as the ESG factor as a novelty are also presented in this section.

The third part is dedicated to Machine Learning (ML) applied to factor investing. This section offers a brief introduction to basic ML concepts, an explanation of some ML techniques and their applications to factor investing, and a final example to illustrate the ideas exposed.

Finally, the conclusions drawn during the study are presented.

1.4. Methodology

In general terms, the methodology followed to develop this article consists of the following steps:

Firstly, a careful literature review has been conducted to better understand the main topic of this thesis: Factor Investing and the possibility of improving the model performance adding ESG criteria as a factor.

Secondly, there have been reviewed some Machine Learning concepts as well as the main algorithms that have been popularly used in the financial area.

Before concluding, a practical example of a Machine Learning algorithm applied to an ESG Factor Investing strategy has been developed.

Finally, after the analysis carried out some conclusions have been reached.

2. FACTOR INVESTING

2.1. Origin

Before factor investing appeared, either active or passive management techniques were used to manage investment portfolios (Elejabeitia, 2018).

On one hand, **active management** is based in the fact that markets are not efficient. This technique focuses on beating the market or a benchmark index by selecting securities that yield the highest return for a certain level of risk. This is done by running an exhaustive analysis and, thus, active management entails high management costs (Lopez, 2022).

On the other hand, **passive management** consists in replicating a benchmark index or market to obtain the exact same returns. This is done by investing in a portfolio

that replicates the index securities and its weights. Therefore, passive management is less costly. This technique is based on the idea that markets are efficient and thus, it is not possible to beat the market as active managers aim to (Lopez, 2022).

Factor investing appears as a midpoint between both techniques and its origin relies on several financial models aimed at optimizing portfolio returns, specifically, the **Modern Portfolio Theory** of Markowitz and the Capital Asset Pricing Model (**CAPM**) (Elejabeitia, 2018).

Modern Portfolio Theory

In 1952 Markowitz completely changed the investment world with the idea of taking risk into account when trying to optimize investments. According to the **Modern Portfolio Theory** that he established, efficient portfolios are those that minimize risk for a certain rentability, or yield the highest possible return for a given level of risk. Therefore, he defined the efficient frontier as the one composed of the portfolios that maximize return for a given level of risk, i.e., the efficient portfolios (Elejabeitia, 2018).

In this model, the risk is measured as the standard deviation while the ***Sharpe ratio*** is used to assess the assets' performance.

$$\boldsymbol{Sharpe\ Ratio} = \frac{(\boldsymbol{Rc} - \boldsymbol{Rf})}{\boldsymbol{\sigma}\ (\mathbf{Rc})}$$

Where:

$$\boldsymbol{Rc} = portfolio\ return$$

$$\boldsymbol{Rf} = Risk - Free\ return$$

$$\boldsymbol{\sigma}\ (\mathbf{Rc}) = Portfolio\ standard\ deviation$$

The *Sharpe ratio* measures the excess return of a certain investment over the risk-free asset, relative to the risk assumed measured by the volatility of the portfolio. Therefore, efficient portfolios are the ones that maximize this ratio (Fernando, 2023).

Capital Asset Pricing Model

Some years later, during the 1960´s decade, William Sharpe, John Lintner, Jan Mossin and Jack Treynor were the main developers of the **CAPM**. This model assumes that markets are efficient, that investors are rational individuals and defines the expected return of a certain asset as a linear function of its *beta* which is a measure of the asset's sensitivity to market movements. *Beta* captures systematic risk, i.e., the portion of risk that cannot be eliminated by diversification.

$$\boldsymbol{\beta} = \frac{\boldsymbol{\sigma cm}}{\boldsymbol{\sigma^2 m}}$$

Where:

$$\boldsymbol{\sigma cm} = \ Asset\ and\ Market$$

$$\boldsymbol{\sigma^2 m} = \ Market\ variance$$

More specifically, the CAPM proposes that the expected returns for all securities can be broken into two components: a risk-free component and a component received for bearing market or systematic risk. Therefore, according to the CAPM the return of a given asset is measured with the following formula:

$$\boldsymbol{Rc} = \ \boldsymbol{Rf} + \ \boldsymbol{\beta} \ \times (\boldsymbol{Rm} - \boldsymbol{Rf})$$

Where:

$$\boldsymbol{Rc} = portfolio\ return$$

$$\boldsymbol{Rf} = Risk - Free\ return$$

$$\boldsymbol{Rm} = Market\ retunr$$

$$(\boldsymbol{Rm} - \boldsymbol{Rf}) = Market\ risk\ premium$$

In this model, Beta is the only risk factor compensated by the market and, as it can be observed in the formula above there´s a positive linear relationship among rentability and risk. Therefore, higher returns imply higher risk.

The CAPM is then the first factor model as it uses beta as the only rentability factor (Elejabeitia, 2018). The idea behind this model is that an investor should be compensated just for the systematic risk faced, i.e., how sensible to the market a security is (measured by beta). However, the investor is not compensated for the idiosyncratic or particular risk of a given asset, as this latter one can be eliminated through the portfolio diversification (Universidad Pontificia Comillas & CFA Institute, 2022).

Both the Modern Portfolio Theory and the CAPM expanded quickly and were widely used in the financial sector. However, as time passed and academic research regarding this topic evolved, many studies that found limitations to these models were published. Concretely, the fact that the CAPM is a one-factor model made it a poor model and further studies showed the existence of new risk factors apart from beta to enhance portfolio returns, for example the Black Scholes Model (Lopez, 2022).

However, the most relevant one regarding the origin of factor investing was the **Fama & French 3 factor Model** published in 1993. Fama & French stated that beta wasn´t significant enough to explain a financial asset rentability. In addition, they proposed new variables that better explained the return of a given asset. As a result, they developed the Three Factor Model according to which a portfolio's or an asset´s returns are given by:

- Its sensitivity to market movements, ie., the CAPM *beta*.
- The assets' sensitivity to its size. This factor was called Small Minus Big (SMB).
- The sensibility to its profitability, measured by the Book-to-Market ratio. This factor was called High minus low (HML) (Muguerza, 2014).

Therefore, this model is represented by the following equation:

$$\boldsymbol{E\ (R_i) = Rf + \ \beta_i * (R_M - Rf) + \ \beta_{SMB} * E(R_{SMB}) + \ \beta_{HML} * E(R_{HML})}$$

Where: $[\boldsymbol{Rf + \beta_i * (R_M - Rf)}]$ is given by the CAPM; $\boldsymbol{E(R_{SMB})}$ represents the expected rentability from the SMB factor; $\boldsymbol{E(R_{HML})}$ is the expected rentability from the HML factor; $\boldsymbol{\beta_{SMB}}$ and $\boldsymbol{\beta_{HML}}$ represent the sensibility of the expected return of asset "**i**": $\boldsymbol{E\ (R_i)}$ to changes in $\boldsymbol{E(R_{SMB})}$ and $\boldsymbol{E(R_{HML})}$, respectively (Lopez, 2022).

Once the Fama & French Model was published, a lot of research on factor investing was run, however, it wasn´t until 2008, after the economic crisis that factor investing was actually implemented in portfolio management. This occurred for the following reasons:

- Firstly, the financial crisis proved that active management techniques weren´t good enough to fight market collapses. As a result, active investors started using passive management techniques. However, even though these latter investment techniques weren´t exposed to the consequences of a bad decision from an active manager, passive investments had a high arbitrage risk while yielding low returns compared to the risk faced. Therefore, investors started using new diversification techniques based on the variables that generated assets' returns. Specifically, investors were looking for investment alternatives that were less costly than passive ones but equally profitable than active investments.
- Secondly, factor investing was widely developed as a result of the popularity of ETFs and new investment alternatives that didn´t condition the portfolio's volatility.
- Thirdly, investors opted for factor investing as a stronger alternative for portfolio diversification rather than traditional portfolio allocation through a wide variety of asset types. This has been demonstrated by various studies, such as the one conducted in 2012 by Antti Ilmanen and Jared Kizer (Ilmanen, et al., 2019).
- Lastly, this investment technique allows investors to get exposure to specific factors, which is currently highly valued by investors given that ESG objectives are very popular in the investment world these days (Elejabeitia, 2018).

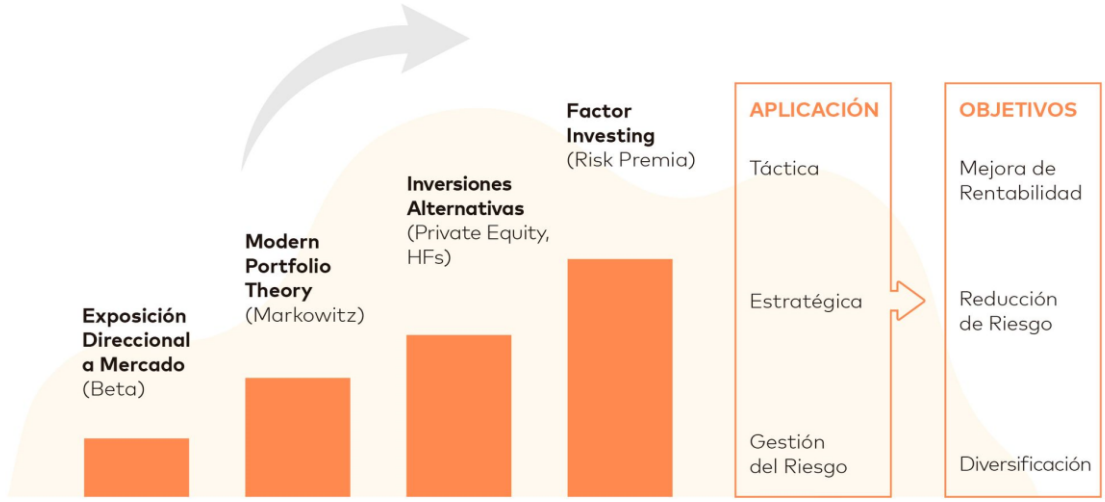

Figure 1 Historical evolution of Investment Techniques.
Source: (Open Bank & Santander Asset Management, 2021).

Figure 1 explains in a very schematic way the evolution of investment techniques and the origin of factor investing.

Having already analyzed the origin of factor investing, we are now going to further study what factor investing exactly is.

2.2. Definition

Factor investing is an investment technique that consists in choosing securities with certain characteristics associated with higher returns (Chen J. , 2020). This strategy is designed to beat the market, obtaining higher returns adjusted to a certain level of risk while maintaining low costs.

The main goal of this technique is to explain which are the drivers of asset prices. The theory behind factor investing is that factors determine the financial performance of firms. These factors could be latent and unobservable, or related to intrinsic features.

Therefore, firms that comply with specific factor requisites should perform better in the future as they have yielded higher risk – adjusted returns over time (Afi - Analistas Financieros Internacionales, 2020).

Moreover, investment strategies based in factors select, weight, and rebalance portfolios favoring these previously mentioned stocks. This is done after a systematic analysis (Nielson, Nielsen, & Barnes, 2016).

As stated by financial analysts from *AFI* (Afi - Analistas Financieros Internacionales, 2020), traditional indexes were constructed based on stock market capitalization so that high capitalization companies have more weighting in the indexes. Therefore, when building portfolios by replicating indexes, stocks would be selected according to their market capitalization.

However, they explain how this phenomenon, i.e., purchasing shares of those big companies, would entail an upward pressure on their price and thus, an increase in their weighting in the index. Consequently, this could lead to market bubbles. As a solution, they mention other possible criteria for stock selection. For example, an index could be constructed picking stocks according to specific characteristics (factors).

2.3. Main factors

The previously mentioned characteristics are called "f**actors**" and there are **two** main **types** of factors: **macroeconomic** factors and **style** factors.

- On one hand, **macroeconomic** factors refer to the systematic sources of risk that affect the whole economy, for example, interest rates, GDP growth or inflation.
- On the other hand, **style factors** are specific stock characteristics that aim to explain excess returns of certain assets, for example, value or volatility. In addition, style factors are the most used ones.

During the last few years, research has been done on a wide variety of new factors, ranging from financial statement´s measures to more technical indicators such as momentum.

In fact, as they mention in (Blitz & Hanauer, 2020), the concept of "factor zoo" appeared to reflect the vast amount of literature on new factors that was being published. However, (Blitz & Hanauer, 2020), found that many of them may be false or "simply lucky findings". Their article explains how the results of each article depend on many aspects and thus, many problems arise when having such a wide range of different studies on the same topic.

Furthermore, in (Blitz, 2020) they propose a homogenization. Specifically, they bundle factors into a small number of groupings, or composite factors.

However, the most popular factors in the current equity markets are:

- ***Value***: this factor captures excess return on stocks that are undervalued relative to their fundamental value. It is typically captured by indicators as the book-to-market ratio or earnings to price ratio.

- ***Volatility***: it refers to stocks with low volatility and a stable beta that is close to one. These stocks yield higher returns. Volatility factor is typically measured by the standard deviation.

- ***Quality:*** it captures excess returns on stocks with some "quality" metrics such as higher margins, higher ROE, or stable earnings.

- ***Size:*** this factor refers to small cap companies that yield higher returns. This factor is measured by market capitalization.

- ***Momentum:*** it captures excess returns on stocks with an upward trend in the past. It is commonly measured by the 12 last month´s relative returns or historical alpha.

As previously mentioned, all these five factors refer to different stock characteristics which are the drivers of stock returns (Bender , Briand, Melas, & Subramanian, 2013). However, it is still not clear whether those **excess returns** come from the **exposure to** certain **factors**, or if it is a consequence of **market inefficiencies** (Lopez, 2022).

Firstly, it is a well-known fact that higher risk implies higher return. Therefore, while **increasing exposure** to certain **factors that entail higher risk**, investors are compensated with higher returns. More specifically, the prime return is to compensate the "systematic risk" that investors cannot avoid by any means. Regarding the previously explained factors, the reasoning would be the following:

- Regarding size, small cap companies are riskier as they are less liquid and frequently have higher credit risk. Moreover, they are often less transparent and not reveal very much information for investors. Consequently, investing in these companies must be compensated with higher returns.
- When it comes to value, investing in undervalued companies is thought to be riskier as they are more sensible to shocks in the economy. Consequently, investors ask for a return premium (Bender , Briand, Melas, & Subramanian, 2013).

Secondly, regarding **market inefficiencies**, from which investors benefit to get higher returns, it can have various causes:

- Market inefficiencies can be a result of investor´s behavioral biases.
  "Behavioral finance" explains that due to cognitive or emotional weaknesses investors act in ways that cause market inefficiencies. For example, it has been studied that people tend to invest in stocks of companies of its own country, being this called "home bias". Moreover, there is proof of other behavioral biases such as overconfidence, or the "loss aversion bias". The latter one explains how investors' concerns about losses on stocks that have performed good in the previous periods tend to be lower as they think the previous successful performance would compensate for the possible upcoming losses (Barberis & Huang, 2001).

- Another possible cause of market inefficiencies is investment constraints coming from industry and regulatory practices. For example, different time horizons or leverage constraints (Heckel, Amghar, Haik, Laplénie, & Leote de Carvalho, 2021).

2.4. ESG Factor Investing

It is a well-known fact that nowadays people are getting more and more aware of the importance of sustainability.

For example, a study from the World Economic Forum (World Economic Forum, 2022) states that "*Gen Z cares about sustainability more than anyone else – and is starting to make others feel the same*".

Furthermore, a joint study from McKinsey and NielsenIQ shows an upward trend of the consumer's purchases of sustainable products reflecting their concern about sustainability (Frey, Bar Am, Doshi, Malik, & Noble, 2023).

Therefore, Environmental, Social and Governance (ESG) criteria and sustainable practices are increasingly evolving and especially after the Paris Agreement was signed and the United Nation´s 2030 Agenda for Sustainable Development was adopted (Naciones Unidas, 2015).

Sustainability can be applied to many areas; however, this investigation is focused on ESG applied to finance. More specifically, this work aims to study the inclusion of ESG criteria in factor investing to possibly yield higher returns.

First of all, to better understand ESG factor investing, I´m going to briefly explain what **ESG criteria** is according to (Ulrich, 2016):

- **Environmental**: this area refers to the contribution and performance of a business regarding environmental challenges, i.e., the measurement of a company´s environmental impact and its management. To do so, specific factors are analyzed; for example, a company´s attitude towards climate change, the use of renewable energies or the management of air and water pollution caused by the operations of the company.

- **Social**: these factors focus on the relationship of a certain company with its social environment. It analyses a wide range of business activities. For example, the company´s treatment to its employees, its policies regarding diversity, salaries, etc.

- **Governance**: corporate governance factors refer to the way a company is managed by its governors, the relationship with shareholders and their rights, and the executive manager´s responsibilities. For example, governance criteria focus on compliance systems, decision-making processes, or the organizational structure.

Even though it is a currently popular research field, studies on this topic started a while time ago. In fact, since the 1970s, a lot of research has been done on the relationship between ESG criteria and Corporate financial performance (CFP). Furthermore, according to (Friede, Busch, & Bassen, 2015), who run an exhaustive study on this area, the relationship between ESG and CFP is very well empirically founded and has been historically proved by many academics to be positive and stable over time.

In line with this previously mentioned positive relationship, sustainable investing is an increasingly popular trend among investors. For example, (Badía, Pina, & Torres, 2019) found that socially responsible investment has notably increased.

The truth is that markets are evolving to adapt to these new sustainable trends, for example, developing new investment vehicles. In fact, in the last few years investments in sustainable funds have increased exponentially (Loscos, 2021).

There are two reasons that explain this phenomenon:

- Firstly, sustainable investing seems to be better in terms of rentability.
- Secondly, the increasing consciousness of sustainability issues is leading investors, and people in general, to develop sustainable activities reflecting its compromise with the planet and its care (Loscos, 2021).

As a result, market valuations are being impacted and traditional investment techniques are evolving. Regarding this change, during the last few years research on ESG factor investing has been done.

Research has shown that ESG criteria has an impact on financial markets. For instance, (Matsumura, Prakash, & Vera-Munoz, 2014) proved how markets positively reacted to carbon emissions disclosure. (Faccini, Matin, & Skiadopoulos, 2021) discovered that financial markets reacted according to expectations on US government decisions regarding climate change fight. In contrast, (Bolton & Kacperczyk, 2021) found that high emission companies yielded higher returns.

As we can see, there are some divergences on whether the impact is positive or negative. According to (Gimeno & González, 2022) this can be possibly explained by the different sample periods and companies.

In spite of this debate, many authors claim that the return of a portfolio can be enhanced by getting exposure to ESG factors combined with traditional risk factors such as momentum, value or size. The aim of this study is to confirm this positive relationship. However, this idea is very general, and the results of incorporating ESG criteria in factor investing depend on how ESG factors are measured as well as which other risk factors are included in the model together with the ESG ones.

For example, (Gimeno & González, 2022) focused on carbon emissions, and thus, the environmental field, to construct an ESG factor. More specifically, in their paper "The role of a green factor in stock prices. When Fama & French go green" they extend previous factor models developed by Carhart and Fama & French by adding the GMP (Green minus polluter) factor. As a result, they propose a model according to which excess returns are explained by the following factors:

- Market factor
- Size: they use the Small Minus Big (SMB) Fama & French factor.
- Growth-value factor: i.e., the High Minus Low (HML) Fama & French factor.
- Momentum: they incorporate Carhart´s Winners Minus Losers (WML) factor.
- Profitability: they use the Robust Minus Weak (RMW) Fama & French factor.
- Investment opportunities: measured with the Conservative Minus Aggressive (CMA) Fama & French factor.
- Carbon emissions

Therefore, their model is given by the following equation:

$$R_{it} - Rf_t = \alpha + \beta(\mathrm{E}[Rm_t] - Rf_t) + \Upsilon_{WML}(W_{MOM,t} - L_{MOM,t}) +$$

$$+\Upsilon_{SMB}(S_{MS,t} - B_{MS,t}) + \Upsilon_{HML}\left(H_{\frac{B}{M},t} - L_{\frac{B}{M},t}\right) +$$

$$+\Upsilon_{RMW}(R_{PROF,t} - W_{PROF,t}) + \Upsilon_{CMA}(C_{INV,t} - A_{INV,t}) +$$

$$+\Upsilon_{GMP}(G_{CO2,t} - P_{CO2,t}) + e_{ei}$$

To develop their new factor, that also proves to be relevant to approximate climate change exposure of firms that are less transparent regarding sustainability related information, they use carbon emissions measured by the ratio on the tons of CO2 equivalent emissions disclosed per million of US dollars in income.

In their analysis they conclude that the Green Factor can explain excess returns even to a higher extent than more traditional factors. Furthermore, they state that the overperformance of greener stocks reflects the increasing consciousness of sustainability in financial markets. Even though regulation on this field has yet to evolve, they highlight the importance of companies evolving towards a more sustainable business model to survive in the near future (Gimeno & González, 2022).

In addition, (Melas, Nagy, & Kulkarni, 2016) confirms the positive impact of ESG integration on both passive and active investment techniques. After analyzing data of the last ten years, they additionally examined the impact of ESG integration on the ability of each strategy to achieve its investment goal. This is done because past returns are no guarantee of future returns.

Factor exposures are proxies of expected returns of factor strategies. Therefore, in their study, they use the percentage reduction in active target factor exposure as a measurement of the effect of ESG limitations on the ex-ante information ratio (IR) of these strategies. Specifically, they evaluate six different strategies.

They found that, in general terms, ESG integration had a modest impact on target factor exposure. Furthermore, the study shows that each strategy is affected differently: Minimum Volatility and Quality strategies suffered a lower impact than Value, Size, Momentum and Yield strategies.

On their behalf, (Fan & Michalski, 2020) examined the impact of integrating ESG criteria to factor investing strategies in the Australian equity market.

In their study, they found that non – ESG screening led to a lower performance, and they explain that this happens because of the lack of interaction between ESG factors and other critical fundamentals.

On the other hand, they demonstrate the outperformance of strategies that combine ESG criteria with other factors such as quality or momentum. Furthermore, they discovered that integrating Environmental, Social and Governmental ratings individually into factors yields higher returns that integrating ESG ratings.

Finally, they explain the better results of ESG integration during adverse market conditions.

Another study regarding ESG factor investing is the one developed by (Roncalli, Le Guenedal, Lepetit, Roncalli, & Sekine, 2020), which basically replicates and confirms the results of the analysis run by (Görgen, et al., 2020).

Specifically, they build a Brown minus Green factor (BMG) but with more basic data than (Görgen, et al., 2020) and add it to the Fama French 1992 model (Fama & French, 1992). Furthermore, in their article they study the carbon risk impact on investment portfolios and analyze how to manage it properly.

3. MACHINE LEARNING APPLIED TO FACTOR INVESTING

3.1. Introduction

What machine learning is?

Machine learning (ML) is a branch of Artificial Intelligence (AI), which is focused on imitating intelligent human behavior through the use of data and algorithms. This concept appeared in the 1950s, and it was defined by Arthur Samuel, an AI pioneer, as "*the field of study that gives computers the ability to learn without explicitly being programmed*" (Brown, 2021) & (IBM, n.d.).

How ML works

Machine Learning starts with data. Programmers gather and prepare the data, selecting a part of the initial dataset to be used to train the ML model. This latter one is called training set.

Once this data preparation step is completed, it´s time to choose which ML model to use. After that, the data is input into the model and it starts the "training phase" of the computer model, during which the model is supposed to find patterns or make predictions.

The data that was held out from the training set, called, test set is afterwards used in the "validation or test phase" to evaluate the performance of the model. The programmer can improve the model at this point to try to get the minimum error and optimize the model performance.

Finally, the last step would be to apply the model to real world datasets.

ML functions

ML models can be used with different purposes, and they can develop three functions. According to (Malone, Rus, & Laubacher, 2020) the function of a Machine Learning system can be:

- **Descriptive:** ML system uses the data to explain what happened.
- **Predictive:** ML system uses the data forecast what will happen.

- **Prescriptive:** ML system will use the data to make suggestions about what actions to take.

ML subcategories

There are three primary ML subcategories:

1) Supervised Machine Learning

Labeled data sets are used to train the model, so that it learns the relationship between input variables and output variables. Then, the model is able to predict the output value for new cases that haven't been used in the learning or training process (Universidad Pontificia Comillas, 2019). Supervised ML is the most used method.

2) Unsupervised Machine Learning

Unsupervised models find regularities in the input data by discovering patterns or trends. The algorithm extracts knowledge from the input data without telling you what to learn (Universidad Pontificia Comillas, 2019).

3) Reinforcement Machine Learning

In Reinforcement Learning models are trained through a trial-and-error process. The learner is a decision-making agent that receives rewards or penalties for the actions it takes when trying to solve a problem. Therefore, the machine should learn the best policy, which is the set of actions that maximize the total reward (Universidad Pontificia Comillas, 2019).

As a brief summary of the three subcategories of Machine Learning, Figure 2 explains what each ML system is used for:

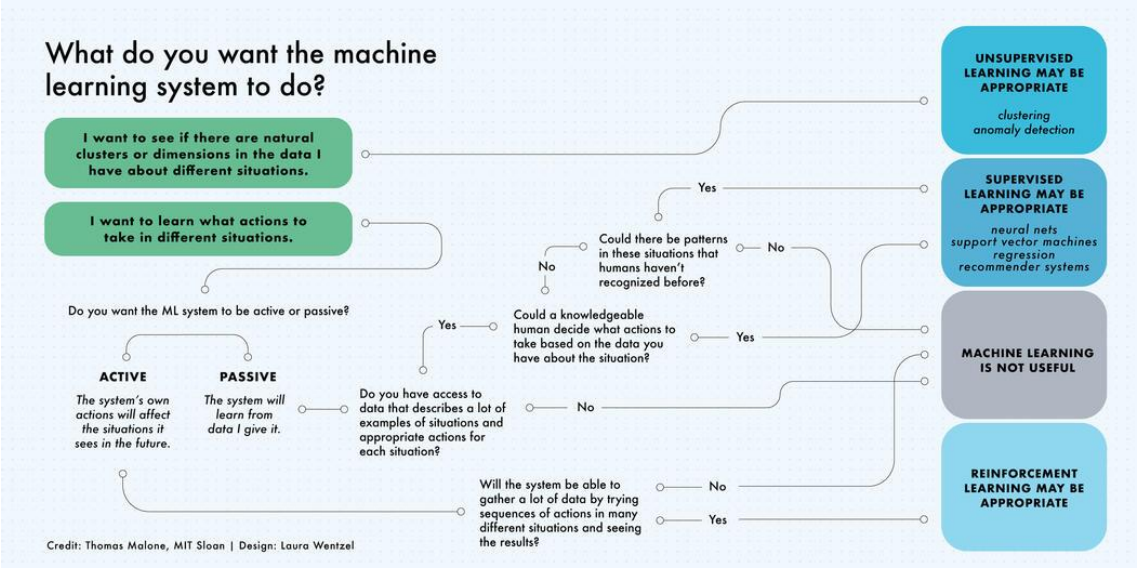

Figure 2 What do you want the Machine Learning system to do?
Source: (Malone, Rus, & Laubacher, 2020).

Pros and Cons of ML

It is a well-known fact that in the last few years the amount of data being generated is increasing exponentially. For example, according to (United Nations, n.d.), 90% of the world's data has been generated in the last two years. Furthermore, they predict a 40% annual increase in the volume of data generated.

Consequently, human processing of all this information is no longer possible. Instead, technological developments such as ML are very helpful to process all this information. Furthermore, ML models can find patterns that are difficult for the human eye to extrapolate as well as identifying complex nonlinear relationships.

However, ML also has some disadvantages that must be taken into account. ML is based on data; therefore, the performance of the models is highly dependent on the quality and amount of information available to train it. Furthermore, interpreting an algorithm's result may sometimes be difficult. It should also be noted that there´s a high level of error susceptibility.

Despite these possible issues that may arise while implementing ML techniques, it is a field that is constantly evolving and improving. ML has a wide variety of applications in real life, and it has proved to be really helpful for humans. This article is focused on ML applied to finance, more specifically, factor investing.

3.2. ML Applications in Factor Investing

The origin of Machine Learning applied to factor investing strategies is based on three main developments: **data availability, computational power,** and **economic framing**.

1) Firstly, the amount of **data** currently available has increased exponentially (United Nations, n.d.). Consequently, data regarding firm specific characteristics is easy to collect. For example, nowadays there´s quarterly or even monthly data over thousands of stocks with information about a wide set of characteristics of each of them. This results in large datasets suitable to be input in ML algorithms. However, it should be noted that chronological depth of the information is a drawback on this aspect, as accounting information is released on a quarterly basis.

2) Secondly, **computational capacity** has evolved both through hardware and software. On one hand, storage and processing speed are no longer technical issues. In addition, open source is now the norm. As a result, ML is currently accessible to everyone.

3) Finally, the **economic groundings** have experienced an evolution as well. In the 1980´s computer scientists and information system experts introduced the first ML applications in finance. A bit after, academics in financial economics exploited them during the 1990´s and hedge funds in the second decade of 2000. It was by that time when nonlinear relationships became popular in the asset

pricing field. All these contributions prepared the path for more force brute methods that have been developed in the last decade.

According to (Coqueret & Guida, 2020), regarding ML applications to factor investing, the model created should make sense economically. Therefore, they just assume that "*future returns depend on firm characteristics*". They state that the relationship between these two aspects varies over time, and it is probably unknown. For (Coqueret & Guida, 2020), this explains why ML could be useful as it can "*detect hidden patterns beyond the documented asset pricing anomalies*". Furthermore, changing market conditions can be faced through the dynamic training of the models.

3.3. ML Algothms

Once analyzed the pillars of ML applications to finance, we should further explain different types of ML algorithms that could be used to develop factor investing strategies to select the one that fits best for this purpose.

First, we should highlight the fact that one of the most important features of factor investing models is that they should explain assets performance through a specific number of characteristics (factors). Therefore, while developing a ML algorithm it should be considered the trade-off between simplicity and informational power, i.e., the algorithm should be simple enough to be legible, but it should include sufficient factors to explain the asset´s performance (Lopez, 2022).

For this purpose, linear and non-linear models should be analyzed to determine which type to use:

1) **Linear models** use a limited number of factors. Therefore, they are very simple, and it is highly probable that they don't explain the portfolio performance, i.e., they have a low informational power.
2) On the other hand, **non-linear models** provide enough information to explain the portfolio performance. However, they also have some limitations: the model is highly dependent on the training set. Furthermore, as the parameter estimation is done through iterative processes, results may sometimes be inaccurate.

Secondly, regarding the three different ML subcategories previously analyzed, supervised methods are more suitable for factor investing applications.

The aim of this article is to confirm that ESG factor investing yields higher returns. Therefore, using supervised learning techniques predictions of portfolio returns will be done.

We are now going to analyze the most useful supervised ML techniques for factor investing applications. Specifically, the following algorithms have been studied for the case: Penalized Regressions, Support Vector Machines (SVMs), Tree-based methods and Neural Networks.

Both penalized regressions and SVMs are going to be briefly explained as they are being progressively substituted by more powerful ML tools as neural networks.

3.3.1. Penalized Regressions

Research and developments on linear model regressions started a while ago. For example, (Chen, Pelger, & Zhu, 2019) already presented the minimum square optimization concept.

The mathematical functioning of the model is as follows; given ***"X"***, a matrix of predictors, the intended output is a linear function of matrix ***"X"*** columns plus an error term. Therefore:

$$\boldsymbol{\epsilon}: \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

Where:

$$\boldsymbol{y} = \textit{Output vector}$$

$$\boldsymbol{X} = \textit{Matrix of predictors}$$

$$\boldsymbol{\epsilon} = \textit{Error}$$

The matrix of predictors would be the different types of factors.

In addition, $\boldsymbol{\beta}$ is the parameter to estimate, and it will be chosen the one that minimizes the error. The minimum error is the one that minimizes the sum of squares of the residuals ($\boldsymbol{L}$):

$$\boldsymbol{L} = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \boldsymbol{\epsilon}\sum_{\boldsymbol{i=1}}^{\boldsymbol{I}}\boldsymbol{\epsilon^2}$$

Where:

$$\boldsymbol{L} = \textit{Sum of squares of the residuals}$$

$$\boldsymbol{I} = \textit{Observations at time } t$$

$\boldsymbol{\epsilon}$ = *Error*

Regarding regularization of linear models, one possible application are penalized regressions. These latter ones aim to improve the robustness of factor – based models by using penalties. The result can then be used to feed into a factor allocation scheme. In fact, (Han, He, Rapach, & Zhou, 2018) & (Rapach & Zhou, 2019) use penalized regressions in their studies. Specifically, they combine forecasts from individual factors to improve stock return prediction.

3.3.2. Support Vector Machines

Support Vector Machines (SVMs) were a very popular ML algorithm since they appeared, and they can be applied to both classification and regression tasks. However, in the last few years, they have been progressively replaced by other much more developed tools, such as Neural Networks, that are gaining popularity (Coqueret & Guida, 2020).

Therefore, SVMs are going to be briefly explained as they aren´t as powerful as Neural Networks or Random Forests when applied to factor investing.

The main goal of SVMs is to build a model that correctly classifies dots in a plane. There are **two possibilities**:

1) When the dots are **linearly separable**, the support vectors are the lines that maximize the distance between the model and the nearest points that are correctly classified by the model. The goal is to build the most robust model, which, among the models that classify correctly, is the one that maximizes the distance between the vectors.
   For example, in Figure 3 the observations located in the dotted lines are the support vectors.

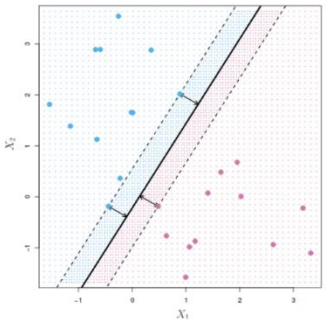

Figure 3 SVM - Case 1
Source: (Universidad Pontificia Comillas, 2019).

2) On the other hand, even though the groups are **not linearly separable** in the original space, they may be linearly separable in a higher dimensional space. Therefore, when the separation of groups is nonlinear, a possible strategy is to expand the dimensions of the original space. To do so, ***kernels*** are used. A Kernel is a function that calculates the dot product of two vectors in a new dimensional space that differs from the original space where they are located.
   An example is illustrated in Figure 4:

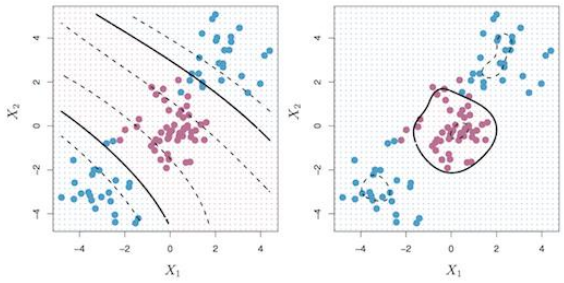

Figure 4 SVM - Case 2
Source: (Universidad Pontificia Comillas, 2019).

3.3.3. Tree – Based Methods

Tree – based methods are supervised ML techniques. These clustering algorithms are simple but very powerful, especially when working with tabular data (Coqueret & Guida, 2020).

Decision trees

Decision trees are the simplest algorithms of this group, and they are aimed at partitioning datasets into homogeneous clusters.

Starting with an **exogenous** variable "**Y**" and **features** "**X**", the process consists in iterative splits of the sample into groups that are as pure as possible in "Y".

This algorithm can be applied to both classification and regression tasks, depending on whether "Y" is categorical or numerical, respectively.

Figure 5 shows an example of how a decision tree works:

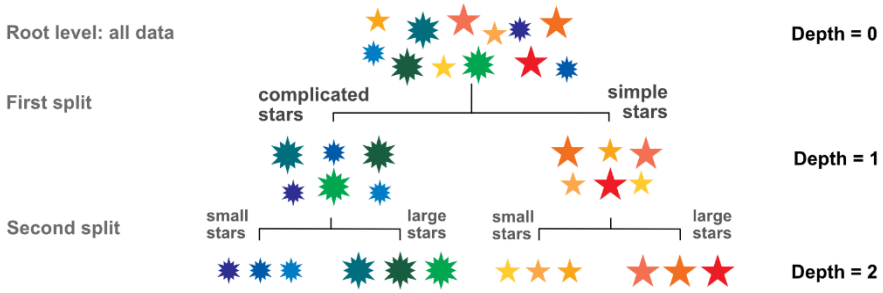

Figure 5 Elementary tree scheme; visualization of the splitting process.
Source: (Coqueret & Guida, 2020).

In this example, the exogenous variable is the "color". Therefore, the goal of the tree is to make various splits of the dataset according to certain variables to homogenize the color within the final sets. For this purpose, two splits are made:

- The first one is made on the feature "complexity" so that stars with five or less points are separated from the rest of the stars.

- The second split is done starting from the two clusters resulting from the first division. Now the division is made according to the "size" of the stars making one group of "large" stars and another one of "small" stars.

The result of these splits are four consistent clusters.

This example can be easily compared to factor investing, in fact it is very explanatory: the color (Y) would be the return of a certain asset. For example, the blue could represent the lowest returns and the red color the highest returns. In addition, the splits would be made according to specific companies features or factors related with the stocks, such as Market Capitalization, or Book to Market ratio. In conclusion, the algorithm would aim to divide a set of companies between the ones that yield higher returns and the companies that don´t.

The mathematical functioning of decision trees is now going to be explained:

Given a sample data set (y_i, x_i) of size *I*, a regression tree will try to divide the data so that the total variance of yi is minimal. This is done in two steps:

- First, the algorithm finds the best partition criteria for each factor x_i^k so that the resulting clusters are as homogenous as possible in y_i.
- Then, it chooses the factor that has achieved the highest level of homogeneity.

Homogeneity in decision trees is linked to the variance. The aim of the decision tree is for the yi inside the subsets to be similar. Therefore, the algorithm tries to minimize the dispersion or variability within each cluster and then sums these two magnitudes. To take into account the relative size of clusters, the algorithm uses "Total variance", i.e., the variance multiplied by the number of elements in each subset.

Mathematically the above is denoted by the following expression:

$$\boldsymbol{V_I^{(k)}(c^{(k)}) = \sum(y_i - m_I(c^k))^2 + \sum(y_i - m_I(c^k))^2}$$

Where:

$$\boldsymbol{V_I^{(k)}} = \textit{Cluster K variance for every observation I}$$

$$\boldsymbol{m_I} = \textit{Average of cluster k}$$

$$\boldsymbol{c^{(k)}} = \textit{Cluster K}$$

In each iteration, the split is made according to the feature that reduces the most the heterogeneity of the resulting clusters, i.e., the one that minimizes the total dispersion over all splits and over all variables. Each split is done starting from the new clusters formed in the previous split.

Decision trees can be cut until the full tree is grown, i.e., each instance belongs to a separate leaf, or the observations of the leaves of a tree can no longer be split according to the defined set of features.

However, this would lead to overfitting problems, i.e., the model would be very accurate with the training data, but useless with out-of-sample data.

It must be said that the first partitions are the most important ones, as they determine the most relevant patterns and agglutinate most of the information. The partitions made at deeper levels of the tree provide less information and explain specific characteristics of each cluster that are of little use to the overall analysis.

To avoid such overfitting problem, it is important to limit the size of the tree. There are several criteria to determine when to stop splitting the data. For example:

- It can be fixed the depth of the tree. The depth is defined as "*the overall maximum number of splits between the root and any leaf of the tree*" (Coqueret & Guida, 2020).
- A minimum gain for each split can be determined so that split is just executed if the reduction of dispersion after the division is above the threshold.
- Impose a minimum amount of data for each terminal node (for each leaf of the tree).
- Setting a condition that the nodes must have a minimum number of observations to make another partition.

Decision trees can be very useful for interpreting relationships between X and Y. However, they can be improved upon the idea of ensembles that combine predicting tools. An example would be **Random Forests**.

Random Forests

Decision trees suffer from high variance. Consequently, there´s **risk of overfitting**, as the algorithm may learn highly idiosyncratic relationships between the data when just one tree is used to make the predictions. Therefore, better predictions can be achieved with ensembles, i.e., by training many trees and combining their classification/regression results (Universidad Pontificia Comillas, 2019).

A **Random Forest** is a particular case of **ensemble** of models. Specifically, the ensemble is done by bagging (Bootstrapped Aggregated Algorithm). The functioning is as follows:

- Firstly, through **bootstrapping**, several random samples are drawn from the original data set with or without replacement. Each tree of the random forest has a different training data set.
- Secondly, a random subset of predictors is used for each split and a tree (unpruned, allowed to grow) is obtained for each data sample.
  The idea is to **end up with a "forest" by fitting *"J"* different trees**.
- Finally, by **combining** the predictions/classifications of the individual trees you **obtain improved predictions**. This is done using the average for regression or the vote (majority) for classification.

Mathematically this is expressed as follows:

$$f^x = \frac{1}{J}\sum_{j=1}^{J} \mathbf{h}_j(x)$$

Where:

$$f^x = \ Random\ Forest\ prediction$$

$$J = \ Total\ number\ of\ trees$$

$$h_j(x) = \ Output\ from\ tree\ "j"\ for\ the\ variable\ "x"$$

The aggregation of all the trained trees via bootstrapping is called **bagging**. This enriches the prediction and makes it more robust. Therefore, it reduces the risk of overfitting.

Furthermore, through bootstrapping the model can identify more general level relationships of variables and features. The random selection of predictors is very important, so that there´s no dominant or influential variable in the first splits of each tree. This allows to improve the predictions as the trees aren´t very similar and are uncorrelated.

In fact, "***m*** ", which is the number of random predictors for each tree, is the main hyperparameter of the Random Forest to be optimized. On the other hand, the number of trees to be added ***"J"*** is not a key factor. Usually many are trained to stabilize the error but without risk of overfitting.

Random Forests return two outputs:

- The first is a set of conditional values.
- The second is the feature importance (FI). Although Random Forest can´t be visualized as decision tress, the FI calculates the relevance of each explanatory variable. Furthermore, this metric shows the predictive power dilution of the model in case one more split was done.

Figure 6 represents an example of how Random Forests work:

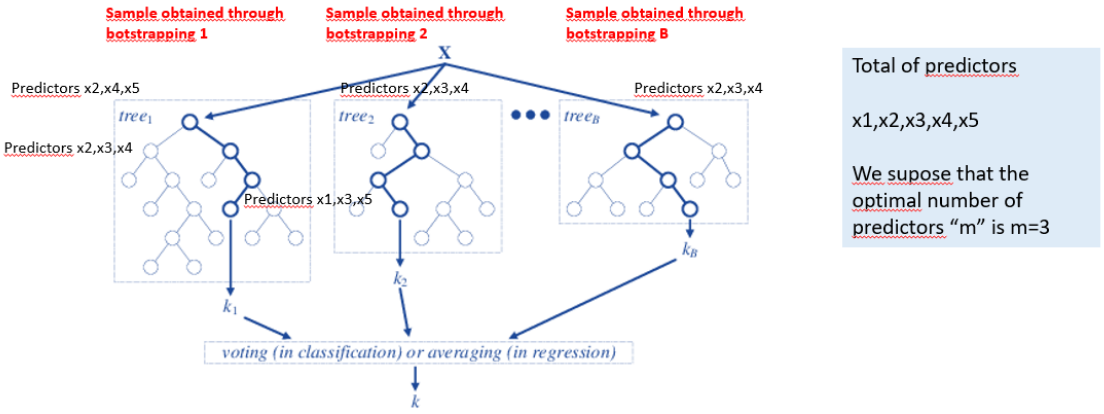

Figure 6 Random Forest.
Source: (Universidad Pontificia Comillas, 2019).

In this case, ***"k"*** would be the Random Forest prediction, which is the weighted average (in case of regression) or majority vote (for classification cases) of the results of each of the ***"B"*** trees.

Applications to factor investing

Regarding factor investing, there have been recently published many articles about the use of tree – based methods in portfolio allocation problems, for example, the ones published by (Ballings, Van den Poel, Hespeels, & Gryp, 2015) & (Patel, Shah, Thakkar, & Kotecha, 2015).

In addition, (Bai, Liu, Yang, & Li, 2019) develop an "S -DT investment strategy" to predict the trend of the closing price of domestic and international stocks. In their study, the discover a high predictive accuracy of this strategy.

On their behalf, (Khaidem, Saha, & Roy Dey, 2016) use a Random Forest to predict the direction of stock market prices. In their study, the algorithm outperformed other algorithms that had been previously used for the same purpose in the literature. In fact, Out of Bag score´s estimations, which are used to validate the model, reflect a successful performance.

3.3.4. Neutral Networks

Neural networks are another rich and powerful algorithm that can be used for factor investing applications. Neural networks can be applied for both regression and classification tasks and, as it is now going to be explained, they are very useful when modeling complex information. They receive this name as they intend to simulate the human brain functioning (Coqueret & Guida, 2020).

In general terms, a neural network is no more than a group of neurons (or nodes) that together form a network. Therefore, neurons are the basic processing unit within a neural network (Universidad Pontificia Comillas, 2019).

To better understand the functioning of this algorithm, it is going to be studied in **three sections**:

- First, it is very important to understand what exactly a **neuron** is and how it works.
- Secondly, it is going to be explained how neurons form **networks** and how do they work.
- Finally, the **training process** of this algorithm will be analyzed.

Neurons

As previously stated, a neuron is the basic processing unit within a neural network. Each neuron has input connections through which they receive external stimuli, which are the input values. With these inputs, the neuron performs an internal calculation and generates an output value (Dot CSV, n.d.). Therefore, a neuron is a mathematical function.

Figure 7 depicts a neuron:

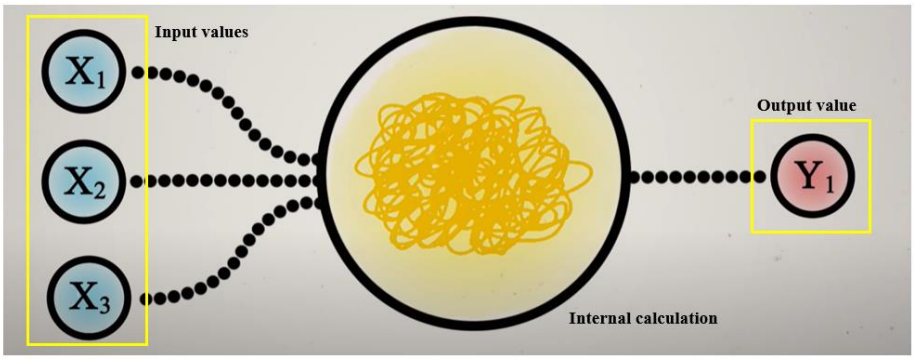

Figure 7 Neuron Basic Scheme.
Source: Own elaboration with data from (Dot CSV, n.d.).

Regarding the **internal calculation**, what each neuron does is to perform a **weighted sum** of all the input values. Each input variable affects the neuron with a specific intensity. In fact, this intensity is given by the weight associated to each connection arriving at the neuron. Furthermore, these weights are the parameters to be estimated in the model, as they can be modified to change the value of the sum positively or negatively. There´s an additional component called ***bias*** and it is an independent term which weight is always one.

At this point, the structure of a neuron would be the as shown in Figure 8:

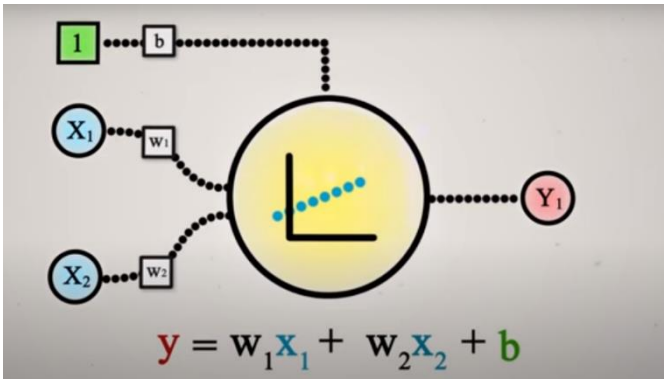

Figure 8 Neuron Basic Scheme.
Source: Own elaboration with data from (Dot CSV, n.d.).

As it can be seen in Figure 8, it is a linear regression. Finally, there is a third component of the neurons that distinguishes them from a simple linear regression equation. This latter one is called the ***Activation Function.*** The activation function is applied to each neuron´s output to add non – linear transformations.

The choice of a particular Activation Function is also an important decision. Figure 9 shows the most common functions:

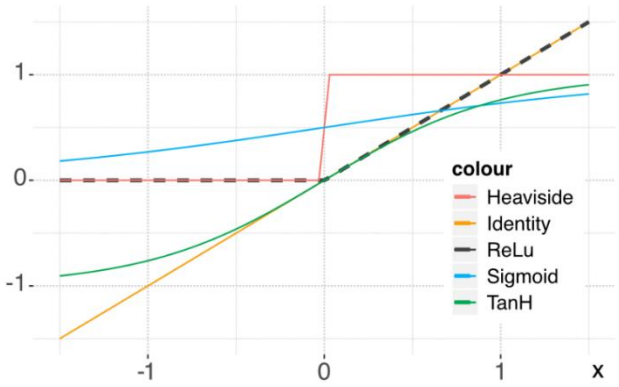

Figure 9 Plot of the most common activation functions.
Source: (Coqueret & Guida, 2020).

Networks

As it has been previously explained, the neurons are grouped to form a network. Specifically, the nodes are organized in different layers. The network is composed of:

- An **input layer**, which collects input values.
- One or more **hidden layers**, in which "internal calculations" are performed.
- An **output layer**, which returns the result.

Therefore, a neural network would look as depicted in Figure 10:

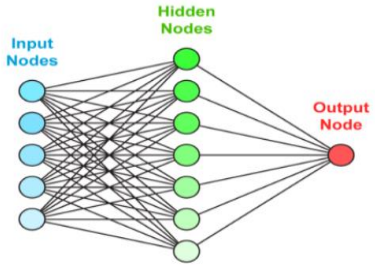

Figure 10 Simplified scheme of a multi-layer perceptron.
Source: (Coqueret & Guida, 2020).

The functioning of the neural network is the following:

- Firstly, once the data is input in the network, the initial linear mapping takes place.

- Then the data is transformed by the activation function. The result of this iteration will be the input for the nodes of the next layer.

- The linear forms are repeated in each layer of the network. Furthermore, each layer is connected to the previous one through the outputs of the preceding nodes, i.e., the output of layer *"l"* is the input of layer "l+1". Those outputs are linear observations to which activation functions have been applied.

- Finally, the results of the network are aggregated into the output layer by the last node. To do so, it employs the weighting mechanism previously explained.

Mathematically, the whole process would be as reflected in Figure 11:

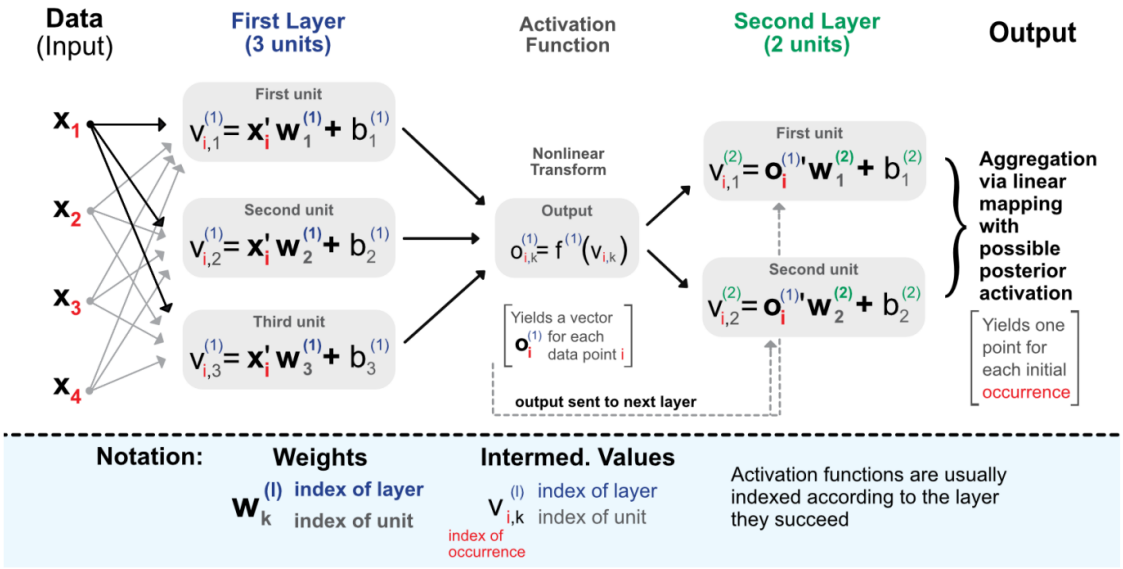

Figure 11 Detailed scheme of a perceptron with 2 intermediate layers.
Source: (Coqueret & Guida, 2020).

When applied to factor investing, the input data *(X)* are the different factor or characteristics that define a specific company.

Training process

Neural networks are trained through the minimization of a loss function subject to some penalization (Dot CSV, n.d.).

Regarding the loss function, the most popular ones are the sum of squared errors for regressions and cross-entropy for classification work. Mathematically, the loss function would be as follows:

$$\mathbf{0} = \sum_{i=1}^{I} \boldsymbol{loss}(\boldsymbol{yi}, \check{\mathrm{y}}\boldsymbol{i}) + \boldsymbol{penalty}$$

Where:

$$\boldsymbol{yi} = \textit{True parameters that the model should return}$$

$$\check{\mathrm{y}}\boldsymbol{i} = \textit{alues obtained by the model}$$

The **penalization** is calculated through the **back – propagation** algorithm. Figure 12 describes the functioning of this algorithm:

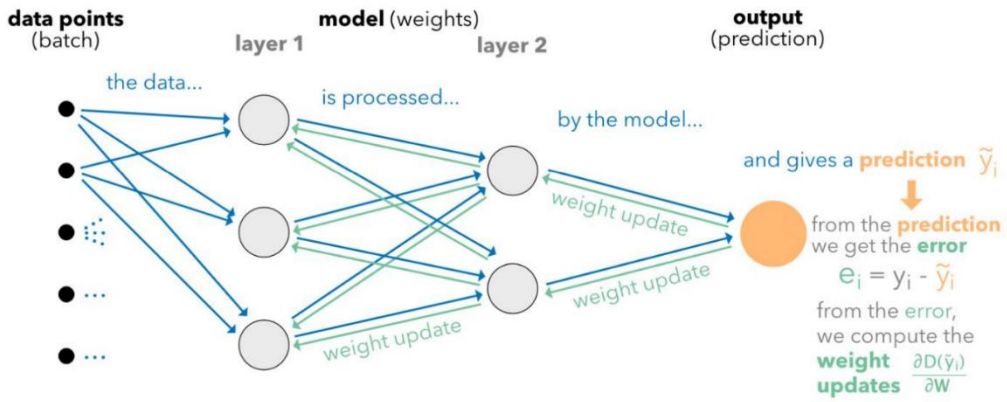

Figure 12 Diagram of back-propagation.
Source: (Coqueret & Guida, 2020).

1) The input data flows through the network nodes until the last layer that gives the output or prediction.
2) From the prediction the information loss function is obtained. In addition, this function gives all the derivatives (weights and biases).
3) These derivatives are then computed backwards starting from the output node. Therefore, each neuron receives the proportional part of the error that it has recorded in the result.

4) In each layer this process is repeated until each neuron of the entire network has received its approximation of the error attribution.

Finally, the model is adjusted, altering the weights and biases for all the neurons of the network, to minimize the error given by the loss function. This is done through the **gradient descent** algorithm. This algorithm determines local minima of a function. Furthermore, the goal is to minimize the error of the model by making the function move in the opposite direction of the gradient, i.e., the slope of the function.

Applications to factor investing

There is a wide range of possible financial applications of neural networks, and, more specifically, to factor investing.

Regarding equities, some popular studies about this topic are now mentioned:

(Feng, He, Polson, & Xu, 2023) build an augmented deep factor model to generate latent factors that best explain the cross-section of stock returns.

Specifically, their goal is to minimize pricing errors and to do so, they first train a neural network using the company´s characteristics as the input; secondly, the model generates risk factors, which are the intermediate features; finally, the outputs of the model are the returns of the securities adjusted to minimize the loss function.

Their model applies the dimension reduction directly to firm characteristics (inputs) instead of factors (intermediate features) or security returns (outputs). Furthermore, their model can learn both nonlinearity and interactions of inputs, which makes it a better choice than Principal Component Analysis techniques. (Feng, He, Polson, & Xu, 2023) model also allows for an unbalanced data structure, for example, individual stock returns.

Finally, the authors find robust statistical and economic evidence in the out-of-sample portfolios and individual stock returns. Furthermore, they achieve through this model significantly better Sharpe ratios.

(Chen, Pelger, & Zhu, 2019) use a complex neural network to estimate the pricing kernel. The network includes a generative adversarial network; thus, it is very useful for portfolio construction as it gives highly important information about the structure of expected stock returns.

In their study, they take advantage of the vast amount of information available and develop an asset pricing model for individual stock returns using complex neural networks. Furthermore, they maintain a fully flexible form and consider time variation.

(Chen, Pelger, & Zhu, 2019) were innovative in their study as they employed the fundamental no – arbitrage condition as a criterion function, so as to build the most informative portfolios within the test set and to abstract the states of the economy from the macroeconomic time series.

The developed asset pricing model was successful as, out-of-sample, it outperforms every benchmark. Performance is measured in terms of Sharpe Ratio,

explained variation and pricing anomalies. In addition, the model identifies the key drivers of asset prices.

3.4. Practical Example

3.4.1. Introduction

After the analysis carried through the literature review, in this section I´m going to perform a simple example of a **Random Forest** applied to a factor investing strategy. The intention is to test the hypothesis of the better performance of an ESG factor model.

For the reader to better understand the methodology followed, it is now going to be introduced the example performed.

A factor model has been developed using 4 factors: Size, Value, Profitability and ESG. These factors have been chosen because, as it has been explained in this article, they have already been used historically to explain stock returns.

However, as the purpose is to test the performance of ESG criteria as an additional factor, the model has been run twice; the first one including the Size, Value and Profitability factors, and the second one adding an additional forth ESG factor.

Once calculated the returns of each company of the stock universe, a portfolio will be constructed. I have chosen to invest in all of the S&P 500 companies; however, the investment decision will be whether to go long or short on each of them. This decision will be given by the binary variable "Strategy".

Furthermore, a Random Forest is applied to run this strategy. Specifically, a classification problem will be performed. The goal of the algorithm is to predict whether to go long or short on each of those companies depending on their individual returns.

Once obtained this information, the returns of the portfolio will be calculated.

After developing the example, the limitations faced will be explained together with some proposals for further studies.

3.4.2. Methodology

3.4.2.1. Data collection

To start the practical exercise the first step was the data collection. Specifically, I gathered the following data from **Bloomberg**: annual data for all the S&P 500 companies and for the last 7 years: 2016 - 2022. The variables of which data was collected were:

- **Price**: This variable has been selected to calculate each stock's returns. As defined by Bloomberg this variable returns "*the last price for the security provided by the exchange*". Given in USD.

- **Price to Earnings (P/E) Ratio**: The P/E ratio information was collected for the value factor calculation. Bloomberg defines it as "*Ratio of the price of a stock and the company's earnings per share".*

- **Market Capitalization**: To calculate the size factor, market capitalization data has been used. As defined in Bloomberg, this variable corresponds to " *Total current market value of all of a company's outstanding shares stated in the pricing currency*". Given in USD.

- **Return on Earnings (ROE):** This variable was chosen to calculate the profitability factor. Bloomberg defines it as "*The measure of a corporation's profitability by revealing how much profit a company generates with the money shareholders have invested, in percentage."*

- **ESG Score**: This has been the variable chosen to calculate the ESG factor. It is defined as "*Bloomberg score evaluating the company's aggregated Environmental, Social and Governance (ESG) performance. The score is based on Bloomberg's view of ESG financial materiality"*. The score's values range from 0 to 10 and 10 is best.

In addition, annual prices for the S&P 500 Index were collected.

The initial data was structured so that there was one dataset for each of the variables mentioned with the dates in the rows and the companies in the columns.

However, the data had to be preprocessed to prepare it to be input into the model. In fact, to better manage the data according to the purpose of the model, it was first restructured so that there was one dataset with information for each year, each of which had the companies in the rows and the variables in the columns. These changes were made using Excel. Once the data was properly structured, the rest of the exercise has been performed in RStudio.

3.4.2.2. Data preparation

After importing the 7 datasets, the same changes were applied to each of them to prepare the final dataset. This latter one will consist of rows with the different observations, corresponding to data from each company at a specific date, and columns with the different factors as well as the dependent variable of the model: "Strategy".

NA´s have been omitted.

At this point the **Factor Calculation** started:

- **Size Factor**: As previously explained, the size factor is aimed to capture the excess returns of smaller companies. As a measure of each company's size Market Capitalization data has been used. Specifically, I have ranked the companies to measure the relative size of each of them to get the Size factor. They are ranked so that the firsts positions correspond to the smallest companies.

- **Value Factor**: This factor captures the excess returns of companies that are undervalued relative to their fundamental value. Among the most popular variables used to calculate this factor, I have selected the P/E ratio. As with the size factor, the companies were ranked according to their P/E ratio to measure this factor. The first positions of the rank are occupied by the companies with the lowest P/E ratios, i.e., the ones considered to be undervalued.

- **Profitability Factor:** The profitability factor aims to capture excess returns of companies with a strong financial performance and profitability. Therefore, this factor has been calculated using the ROE. The methodology has been the same as with the previous factor but with a small difference.

  This factor was developed upon the idea that more profitable companies yield higher returns, therefore, the companies with higher ROE's should have a better profitability factor. Consequently, the rank has been adjusted in this case so that the first positions are occupied by the strongest companies, in terms of financial performance, i.e., the ones with higher ROE's.

- **ESG Factor:** In general terms, this factor aims to capture excess returns of best valued companies according to ESG criteria. This factor is based on recent studies that have shown that not only, ESG criteria is becoming more relevant in the current investment world, but it has proved to well explain stock returns.

  As ESG entail a wide range of aspects, there are many different ways to calculate this factor, as it can be focused on one of the three ESG pillars, i.e., Environmental, Social or Governmental, or it can capture a global measure of all the three pillars.
  In this example, the ESG score has been the selected variable to measure this factor. Since the data collected from Bloomberg gives a higher score to companies that perform better in terms of ESG criteria, the rank has been adjusted as in the profitability factor so that the companies with higher ESG scores are best ranked.

After calculating the factors, the final dataset was completed adding the **dependent variable** ("**Strategy**").

Before explaining the variable development process, I want to emphasize the fact that in this example the investment decision isn´t selecting which companies add to the portfolio, but instead decide whether to go long or short on each stock assuming the portfolio contains all the companies of the universe, i.e., all the S&P 500 companies.

Therefore, the logic to construct this variable is that an investor should go long on the companies that yield higher returns and short on the ones that yield lower returns. To do so, I created this binary variable, setting the cutoff in the median of the companies so that the investor would go long (1) on 50% of the companies, the ones with the highest returns, and short (-1) on the other 50% which are the ones that yield the lowest returns.

Once repeated all these steps for each date the "**Final Dataset**" is prepared to be input into the Random Forest. Its column contain data for all the factors calculated ("Size", "Value", "Profitability" and "ESG") as well as the dependent variable ("Strategy"). The rows correspond to the different observations, i.e., each of the S&P 500 companies at each point in time.

The last step of this data preparation process was the split between the train and the test set. Specifically, 5 years data have been used for the training set and 2 years data for testing the model.

3.4.2.3. The Model

Once the data was prepared, the model has been built. As previously explained the model is developed to determine whether to go long or short on a specific stock based on 4 factors that describe each company.

Two models have been developed "***rf_model_1***" and "***rf_model_2***":

1) ***rf_model_1***": The first model uses the Size Factor, Value Factor and Profitability Factor as the predictors.
   Regarding the parameters of the Random Forest, I have set the most important ones: the number of trees to grow ("*ntree*") and the number of variables randomly sampled at each split ("*mtry*").
   In this first model I have selected 100 trees as the dataset is small. On the other hand, "*mtry*" recommended value is usually calculated as the $\sqrt{P}$ for classification problems, being p the total number of variables of the model, 3 in this example. (León, 2018) Therefore, the "*mtry*" parameter has been set to 2.

2) "***rf_model_2***": The second model uses the Size Factor, Value Factor, Profitability Factor and an additional ESG Factor as the predictors.
   Regarding the parameters of the Random Forest, I have set again both *"ntree"* and *"mtry"*. In this second model I have selected 150 trees, since there's one more variable, however, I didn´t set it higher as the dataset is still small. On the other hand, the "*mtry*" parameter has been set to 2 since there are 4 variables.

Although these are the final models, several trials have been done with both models to adjust the values of the parameters to try to optimize the model.

3.4.3. Results

The performance of the model is going to be analyzed using the most popular performance metrics for classification problems. To do so the "*confusionMatrix*" function of R is used.

The models´ performance has been analyzed using the following metrics:

- **Confusion Matrix**: it consists in a cross – tabulation of the predicted and real or observed classes.

- **Accuracy**: this metric reflects the global precision of the model. It is calculated as the proportion of correctly classified predictions over the total number of predictions. Therefore, the higher this metric is the better the model is.

- **Recall:** it is a measure of the proportion of the positive cases properly classified. It is calculated as the true positive cases, divided by the sum of true positive plus false negative cases.

- **Specificity:** this metric reflects the proportion of negative cases properly classified. It is calculated as the true negative cases, divided by the sum of true negative plus false positive cases.

- **F1 Score:** it is a combined measure of the precision and the recall measures, calculated as the harmonic mean of precision and recall. Therefore, it takes into account both false positive and false negative cases.

- **(Receiver Operating Characteristic) ROC curve:** It is graph that shows performance of a classification model. It is obtained by representing the True Positive Rate in ordinates and False Positive Rate in abscissae.

Model 1 performance

After running the "*confusionMatrix*" function for the first model the following results are obtained:

- **Confusion Matrix**:

```
          Reference
Prediction  -1   1
        -1 201 186
        1  186 201
```

- **Accuracy**: 0,531

- **Recall:** 0,5194

- **Specificity:** 0,5194

- **F1 Score:** 0,5194

- **(Receiver Operating Characteristic) ROC curve:**

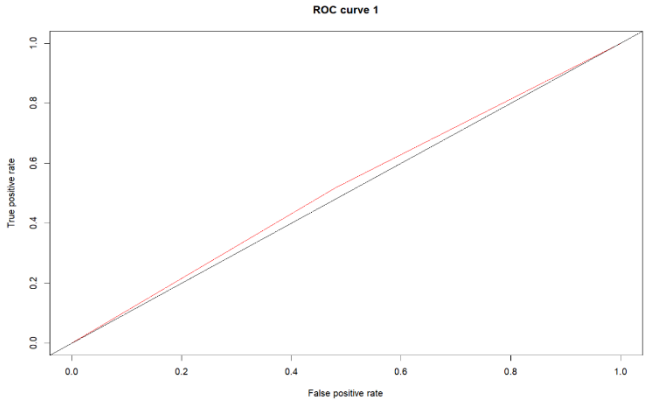

Figure 13 Roc Curve Model 1.
Source: own elaboration with RStudio.

Model 2 performance

After running the "*confusionMatrix*" function for the first model the following results are obtained:

- **Confusion Matrix**:

```
          Reference
Prediction  -1   1
        -1 198 174
        1  189 213
```

- **Accuracy**: 0,531
- **Recall:** 0,5116
- **Specificity:** 0,55
- **F1 Score:** 0,522
- **(Receiver Operating Characteristic) ROC curve:**

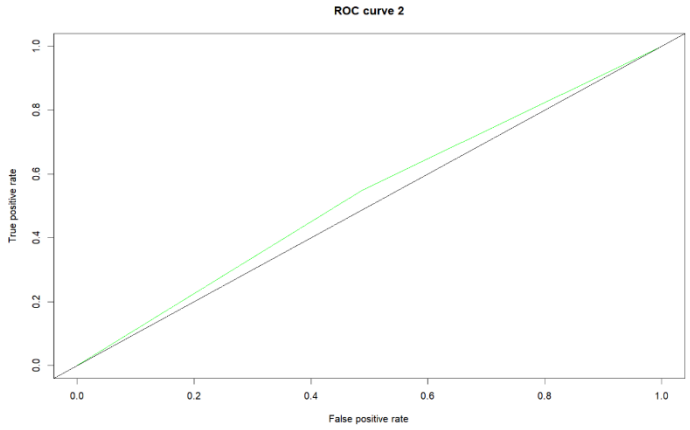

Figure 14 Roc Curve Model 2.
Source: own elaboration with RStudio.

These results reflect a bad predictive performance of both models. However, as I would explain more deeply in the "Limitations" section, these results are highly probably due to the lack of data.

Despite the results, I consider interesting comparing the performance of both models as the main purpose of this example was to proof the idea of the better performance of an ESG Factor investing model.

As it can be seen in the results obtained, both models have a similar performance. However, the "***rf_model_2***" has a higher accuracy than the "***rf_model_1***". Regarding the other metrics evaluated, "***rf_model_2***" has a higher value for the specificity metric while the recall metric is higher for "***rf_model_1".*** Furthermore, when the ROC curves are plotted, as it can be seen in *Figure 13* and *Figure 14*, the "***rf_model_2***" ROC curve shows a better performance than the "***rf_model_1"*** ROC curve.

Despite these latter results, I consider more relevant the accuracy results as this measure captures the global precision of the model, while the others just focus on true positive or true negative cases.

In this example there's no "positive" or "negative" choice. Therefore, given the similar performance results for both models, I believe that accuracy is more important than recall or specificity when it comes to comparison in this example.

Taking the accuracy results into account, the ESG model's performance is better, and thus, the hypothesis would be confirmed. However, this may be due to many possible reasons and not exactly for the high explanatory power of the ESG factor as it has been studied in this article.

For example, the fact that the second model has one more variable could be one of the reasons why it performs better. Furthermore, when analyzing the feature's importance of each model, the ESG variable doesn't appear to be that much relevant.

3.4.4. Limitations

In this section I would like to explain the limitations faced during this study and propose further investigations on this topic.

The main constraint has been the lack of data available, and this has led to several shortcomings during the study. As it has been previously explained in this article, the evolution towards ESG criteria applied to finance and, in general, its increasing importance worldwide, is recent and thus, the world and, in this specific case, the financial sector, is still adapting to this new circumstance. Consequently, data regarding ESG criteria is very scarce.

In this concrete example, ESG data was available just in annual frequency, while the usual frequency for this analysis would have been monthly or even daily data. Furthermore, the first year for which this data was available was 2016 and thus, just 7 years of data could be used. This data scarcity has led to the following problems:

- Due to the annual frequency of the data, the Market Factor hasn't been included in the model. This happened because the market beta couldn't be calculated as having data for 1 point in time for each year it wasn't possible to make the linear regression of the market's return and the company's return.

- In addition, Random Forests and, in general, ML techniques are used to optimize the management and analysis of high amounts of data; however, when applied to small datasets algorithms may not be properly trained and, thus, lead to poor predictive performance as it has been the case in this example.

Considering these limitations, the following solutions are proposed for further studies:

- This is a simple example aimed at testing the performance of an ESG factor model and thus, few variables have been used, focusing more on the impact of this new factor. However, the idea behind the model proposed could be further developed adding more factors to try to enhance the model's performance.

- Furthermore, there are several variables that could be used to calculate the ESG Factor and, depending on the characteristics of the companies and the specific goals of the investor other more specific variables could have been chosen. For example, CO2 emissions that focuses on the environmental pillar, a Gender Diversity Index, focused on the social pillar, or compliance programs that evaluate the implementation and effectiveness of internal regulatory and ethical goals, which is a measure of the governance pillar. Not only these more specific variables could be tested but also combinations of them.

- An additional fact to consider is the possibility of a long – term performance analysis that could be done in the future when more historical ESG data will be available. By extending the study to a longer time frame, the model could capture the impact of ESG factors during different economic conditions, as it has already been done with other factors.

Moreover, this longer time horizon allows the investor for a deeper evaluation of the model's ability to identify companies with long-term value creation potential in terms of sustainability. Most companies have just started their evolution towards ESG integration, so it is probable that the ESG companies' performance will improve over time and, thus, current historical data on this field is not that much illustrative of the future's performance.

- Finally, another possible approach for the study could be to take into account different ESG rating methodologies as each rating agency follows a specific criterion while making ESG ratings. Therefore, it could be interesting to compare ESG ratings' results from different rating agencies and this could help identify the most relevant and reliable ESG factor according to each investment decision.

4. CONCLUSIONS

In this article the main point studied has been the impact of adding ESG criteria as an additional factor in a factor investing strategy.

The first point reviewed has been the origin of this investment strategy to better understand what factor investing is. As it has been studied in this article, Factor Investing is an investment technique that consists in choosing securities that have certain characteristics that are associated with higher returns. These strategies are designed for beating the market, achieving higher returns or reducing risk.

Some popular factors include value (investing in undervalued assets), momentum (capitalizing on recent strong performance), size (investing in smaller companies), quality (seeking financially stable companies), and low volatility (investing in assets with lower price fluctuations).

Once analyzed the concept of Factor investing, a new factor has been reviewed, the ESG factor. Environmental, Social, and Governance criteria have recently gained relevance worldwide, and, in this specific context, in factor investing strategies.

ESG factor investing consists in evaluating companies based on their sustainability practices, social impact, and governance standards. The idea behind it is that companies with strong ESG performance should perform better and thus, yield higher returns. Therefore, adding ESG factors should enhance the risk-return profile of factor investment strategies.

Once reviewed the main topics of the study, ML basic concepts are explained. Additionally, it is studied the possible application of ML algorithms to optimize Factor Investing Strategies as it allows investors to process vast amounts of data, identify meaningful patterns, and make data-driven investment decisions.

Finally, a simple example has been performed to test the performance of ESG criteria as an additional factor. In the example a factor model has been developed using 4 factors: Size, Value, Profitability and ESG. Furthermore, a Random Forest is applied to run this strategy. Specifically, it has been performed a classification problem to predict

whether to go long or short on each of the S&P 500 companies depending on their individual returns.

In the study there were strong limitations regarding data availability, however, the results obtained are in line with the conclusions drawn from the literature review and the addition of the ESG factor improves the model's performance.

Once finalized the analysis I have drawn the following conclusion:

As well as the world is still adapting to new issues and necessities regarding ESG aspects, the financial sector and, more specifically, Factor Investing, still has to evolve too.

I strongly believe that ESG applied to finance has a lot of potential value creation considering the increasing importance that people are giving to this field. Furthermore, investors are already recognizing the importance of considering ESG factors in their decision-making process. By doing so they can align their investment decisions with their values and contribute to a better future as well as benefiting from an improved risk management and better long – term financial performance.

In my opinion, from a more global point of view, the future of finance is going to evolve towards ESG integration with continued developments in ESG data, metrics, and analytical tools. This would be a result of the evolving needs and preferences of investors and society as a whole.

Furthermore, regarding Machine Learning techniques, I think they also have a growth potential. Even though they are already considered valuable when applied to factor investing, its usefulness will increase as well, in line with the growing volume of ESG data availability. As consciousness regarding ESG criteria continues to grow, more data is being generated and collected, such as carbon emissions, labor practices or corporate governance, among others.

The increasing volume of ESG data, together with the power of Machine Learning techniques will result in the development towards ESG integration into investment strategies, resulting in a more complete and accurate assessment of the financial performance and risks associated with ESG factors.

5. APPENDIX

```
#---TFG ANALYTICS---


# I have an Excel with data for 7 years for each of the variables I need to calculate the factors I´ll be using for my factor investing strategy.


# First, I´ll calculate the value of my factors for each year to build the dataset I´ll imput to my Random Forest.
```

```
# Import The libraries I´ll be using


library(tidyverse)

library(randomForest)

library(quantmod)

library(readxl)

library(caret)

library(ROCR)

library(pROC)


# Set working directory

setwd("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS")


# Import data for all S&P500 companies for 7 years

SP500_2016 <- read_excel("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS/S&P-TFG-2016.xlsx", col_types = c("date", "text", "numeric",  "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric" ))

SP500_2017 <- read_excel("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS/S&P-TFG-2017.xlsx", col_types = c("date", "text", "numeric",  "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric" ))

SP500_2018 <- read_excel("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS/S&P-TFG-2018.xlsx", col_types = c("date", "text", "numeric",  "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric" ))

SP500_2019 <- read_excel("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS/S&P-TFG-2019.xlsx", col_types = c("date", "text", "numeric",  "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric" ))

SP500_2020 <- read_excel("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS/S&P-TFG-2020.xlsx", col_types = c("date", "text", "numeric",  "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric" ))

SP500_2021 <- read_excel("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS/S&P-TFG-2021.xlsx", col_types = c("date",
```

```
"text", "numeric",  "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric" ))

SP500_2022 <- read_excel("C:/Users/loret/OneDrive - Universidad Pontificia Comillas/icade/5º
E2-ANALYTICS/TFG/ANALYTICS/BASE DE DATOS/S&P-TFG-2022.xlsx", col_types = c("date",
"text", "numeric",  "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric" ))


#----------DATA PREPARATION----------------------------


#Use a loop to calculate the factors for each year


# Create a vector with the datasets for each year's data

dataset_names <- c("SP500_2016", "SP500_2017", "SP500_2018", "SP500_2019",
"SP500_2020", "SP500_2021", "SP500_2022")


#Create the list where I´ll store my final datasets (with the factors and the variable Y)


Final_datasets <- list()


#Loop


for (i in dataset_names) {


  dataset1 <- get(i)


  #NA´s treatment

  dataset <- na.omit(dataset1)




  #--Factors Calculation--


  #SIZE FACTOR

  #Rank the companies based on their market capitalization
```

```
dataset$SIZE_FACTOR <- rank( dataset$`MARKET CAP`)


#VALUE FACTOR
#Rank the companies based on their P/E ratio

dataset$VALUE_FACTOR <- rank( dataset$`P/E RAT`)


#PROFITABILITY FACTOR
#Rank the companies based on their ROE

dataset$PROFITABILITY_FACTOR <- rank( dataset$ROE*(-1))


#ESG FACTOR
#Rank the companies based on their ESG SCORE

dataset$ESG_FACTOR <- rank( dataset$`ESG SCORE`*(-1))

#--Variable Y Calculation--

#VARIABLE Y
#Rank based on each company's returns

dataset$ret_rank <- rank(dataset$RETURNS)

#Create the "Long"(1) "Short" (-1) variable (Y)
#Cutoff, median of companie´s ranks
```

```
  num_companies <- nrow(dataset)
  threshold <- round(num_companies / 2)
  dataset$Strategy <- ifelse(dataset$ret_rank <= threshold, -1, 1)

  #Store the final results for each dataset

  Final_datasets[[i]] <- dataset[, c(1,2,15,16,17,18, 20)]
}

view(Final_datasets$SP500_2016)

#In the list Final_datasets I have the datasets I´ll be using in my model:
#Each of them has the independent variables (x: the factors of the model) and the dependent variable (y: whether I´ll go long or short in each stock)

#SPLIT Train vs. Test

SP500_16 <- Final_datasets$SP500_2016
SP500_17 <- Final_datasets$SP500_2017
SP500_18 <- Final_datasets$SP500_2018
SP500_19 <- Final_datasets$SP500_2019
SP500_20 <- Final_datasets$SP500_2020
SP500_21 <- Final_datasets$SP500_2021
SP500_22 <- Final_datasets$SP500_2022

# 5 years data for train
train <- rbind(SP500_16, SP500_17, SP500_18, SP500_19, SP500_20)
train$Strategy<-as.factor(train$Strategy)

# 2 years data for test
test <- rbind(SP500_21, SP500_22)
```

```
test$Strategy<-as.factor(test$Strategy)


#------------------THE MODELS-------------------------

set.seed(123)

#MODEL 1- WITHOUT ESG


#Create the Random Forest Model

rf_model_1 <- randomForest(Strategy ~ SIZE_FACTOR + VALUE_FACTOR +
PROFITABILITY_FACTOR , data = train, ntree = 100, mtry= 2, type = "classification")


#Making predictions with the model

predictions_1 <- predict(rf_model_1, newdata = test)


#Feature Importance

feature_importance_1 <- importance(rf_model_1)



#MODEL 2- WITH ESG FACTOR


#Create the Random Forest Model

rf_model_2 <- randomForest(Strategy ~ SIZE_FACTOR + VALUE_FACTOR +
PROFITABILITY_FACTOR + ESG_FACTOR , data = train, ntree = 150, mtry= 2, type =
"classification")


#Making predictions with the model

predictions_2 <- predict(rf_model_2, newdata = test)


#Feature Importance

feature_importance_2 <- importance(rf_model_2)
```

```
#------------------PERFORMANCE EVALUATION-------------------------

#Model 1
perform_1 <- confusionMatrix(predictions_1, test$Strategy)

#Model 2
perform_2 <- confusionMatrix(predictions_2, test$Strategy)


#CONFUSION MATRIX

#Model 1
Conf_mat_1 <- perform_1$table

#Model 2
Conf_mat_2 <- perform_2$table


#ACCURACY

#Model 1
Ac_1 <- perform_1$overall[1]

#Model 2
Ac_2 <- perform_2$overall[1]

#RECALL

#Model 1
Rec_1 <- perform_1$byClass[6]
```

```
#Model 2
Rec_2 <- perform_2$byClass[6]


#SPECIFITY

#Model 1
Spec_1 <- perform_1$byClass[2]

#Model 2
Spec_2 <- perform_2$byClass[2]


#F1 SCORE

#Model 1
F1_1 <- perform_1$byClass[7]

#Model 2
F1_2 <- perform_2$byClass[7]


#ROC CURVES

#Model 1
predictions_1 <- as.numeric(predictions_1)
test$Strategy <- as.numeric(test$Strategy)

pred_1 <- prediction (predictions_1, test$Strategy)
perf_1 <- performance(pred_1,"tpr","fpr")
```

```
plot(perf_1, col = "red", main = "ROC curve 1")

abline(a=0, b= 1)



#Model 2

predictions_2 <- as.numeric(predictions_2)

test$Strategy <- as.numeric(test$Strategy)


pred_2 <- prediction (predictions_2, test$Strategy)

perf_2 <- performance(pred_2,"tpr","fpr")



plot(perf_2, col = "green", main = "ROC curve 2")

abline(a=0, b= 1)
```

6. BIBLIOGRAFÍA

Afi - Analistas Financieros Internacionales. (2020, Abril). Factor Investing & ETF: Gestión por Factores para una inversión más eficiente. *Factor Investing*.

Badía, G., Pina, V., & Torres, L. (2019, April). *Financial Performance of Government Bond Portfolios Based on Environmental, Social and Governance Criteria.* Retrieved from MDPI - Sustainability: https://www.researchgate.net/publication/332771345_Financial_Performance_of_Government_Bond_Portfolios_Based_on_Environmental_Social_and_Governance_Criteria

Bai, M., Liu, X., Yang, K., & Li, Y. (2019). Stock Investment Strategy Based on Decision Tree. *IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, (pp. 151-155).

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). "Evaluating Multiple Classifiers for Stock Price Direction Prediction.". *Expert Systems with Applications 42 (20)*, 7046–56.

Barberis, N., & Huang, M. (2001). Mental Accounting, Loss Aversion and Individual Stock Returns. *Journal of Finance 56*, 1247 - 1292.

Bender , J., Briand, R., Melas, D., & Subramanian, R. A. (2013, December). *Fundations of Factor Investing.* Retrieved from MSCI - Research Insight : https://www.msci.com/documents/1296102/1336482/Foundations_of_Factor_Investing.pdf/004e02ad-6f98-4730-90e0-ea14515ff3dc

Blitz, D. (2020, March). *Cómo orientarse en el "zoo de los factores".* Retrieved from Robeco : https://www.robeco.com/es-latam/vision-de-mercado/2020/03/como-orientarse-en-el-zoo-de-los-factores

Blitz, D., & Hanauer, M. (2020, March). *Factor investing. How many factors are there? Or how to navigate the 'factor zoo'.* Retrieved from Robeco: https://www.robeco.com/docm/docu-202003-how-to-navigate-the-factor-zoo-us.pdf

Blitz, D., & Hanauer, M. (2020, March). *Factor investing. How many factors are there? Or how to navigate the 'factor zoo'.* Retrieved from Robeco: https://www.robeco.com/docm/docu-202003-how-to-navigate-the-factor-zoo-us.pdf

Bolton , P., & Kacperczyk, M. (2021). DO INVESTORS CARE ABOUT CARBON RISK? *Journal of Financial Economics*, 517 - 549.

Brown, S. (2021, April). *Machine Learning, Explained.* Retrieved from MIT - Management Sloan School: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Carmignac. (n.d.). *Carmignac - Sustainable Investments*. Retrieved from Carmignac: https://www.carmignac.es/es_ES/inversion-sostenible/fondos-sostenibles

Chen, J. (2020, 10). *Factor Investing*. Retrieved from Investopedia: https://www.investopedia.com/terms/f/factor-investing.asp

Chen, L., Pelger, M., & Zhu, J. (2019, April). *Deep Learning in Asset Pricing.* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350138

Coqueret, G., & Guida, T. (2020). *Machine Learning for Factor Investing.* Chapman & Hall.

Dot CSV. (n.d.). *Artificial Intelligence - Youtube Chanel*. Retrieved from https://www.youtube.com/@DotCSV/about

Elejabeitia, P. G. (2018, Junio). *GESTIÓN DE CARTERAS. GESTIÓN ACTIVA VS. GESTIÓN PASIVA.* Retrieved from Repositorio Universidad Pontificia Comillas: https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/18660/TFG-%20Gomendio%20de%20Elejabeitia%2C%20Pablo.pdf?sequence=1&isAllowed=y

Faccini, R., Matin, R., & Skiadopoulos, G. (2021, March). *Dissecting Climate Risks: Are they Reflected in Stock Prices?* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3795964

Fama, E., & French, K. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance, Vol XLVII N2*.

Fan, J., & Michalski, L. (2020). Sustainable factor investing: Where doing well meets doing good. *International Review of Economics and Finance 70*, 230 - 256.

Feng, G., He, J., Polson, N., & Xu, J. (2023). *Deep Learning in CharacteristicsSorted Factor Models.* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3243683

Fernando, J. (2023, Mayo 11). *Sharpe Ratio Formula and Definition With Examples*. Retrieved from Investopedia: https://www.investopedia.com/terms/s/sharperatio.asp

Frey, S., Bar Am, J., Doshi, V., Malik, A., & Noble, S. (2023, February). *Los consumidores se preocupan por la sostenibilidad y lo respaldan con sus billeteras.* Retrieved from McKinsey & Company: https://www.mckinsey.com/featured-insights/destacados/los-consumidores-se-preocupan-por-la-sostenibilidad-y-lo-respaldan-con-sus-billeteras/es

Friede, G., Busch, T., & Bassen, A. (2015). ESG and Financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 210 - 233.

Gimeno, R., & González, C. (2022, March). *THE ROLE OF A GREEN FACTOR IN STOCK PRICES. WHEN FAMA & FRENCH GO GREEN.* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064848

Görgen, M., Jacob, A., Nerlinger, M., Riordan, R., Rohleder, M., & Wilkens, M. (2020, August). *Carbon Risk.* Retrieved from SSRN Papers: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2930897

Han, Y., He, A., Rapach, D., & Zhou, G. (2018, June). *Cross-Sectional Expected Returns: New Fama-MacBeth Regressions in the Era of Machine Learning.* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3185335

Heckel, T., Amghar, Z., Haik, I., Laplénie, O., & Leote de Carvalho, R. (2021, February). *Out-Performing Corporate Bonds Indices With Factor Investing.* Retrieved from SSRN: https://ssrn.com/abstract=3697872

IBM. (n.d.). *What is Machine Learning?* Retrieved from IBM: https://www.ibm.com/topics/machine-learning

Ilmanen, Antti, Ronen, I., Tobias J, M., Thapar, A., & Wang, F. (2019). Factor Premia and Factor Timing: A Century of Evidence. *SSRN Working Paper 3400998*.

Khaidem, L., Saha, S., & Roy Dey, S. (2016). Predicting the direction of stock market prices using random forest. *AppliedMathematicalFinance Vol.00,No.00*.

León, M. G. (2018). *Análisis de sensibilidad mediante Random Forest.* Obtenido de Repositorio Universidad Politécnica de Madrid : https://oa.upm.es/53368/1/TFG_MARTA_GARCIA_RUIZ_DE_LEON.pdf

Lopez, J. P. (2022, Junio). *APLICACIÓN DEL MACHINE LEARNING AL FACTOR INVESTING EN RENTA FIJA CORPORATIVA.* Retrieved from Repositorio Universidad Pontificia Comillas: https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/56412/TFG%20-%20Periel%20Lopez%2C%20Jose.pdf?sequence=-1&isAllowed=y

Loscos, A. G. (2021, Marzo). *LA INTEGRACIÓN DE LOS ESG EN LOS PROCESOS DE INVERSIÓN: UNA PROPUESTA METODOLÓGICA PARA LA TOMA DE DECISIONES.* Retrieved from Repositorio Universidad Pontificia Comillas: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.comillas.edu/images/catedras/catedra-impacto-social/TFG%20ANA%20Gutierrez-Premio.pdf

Malone, T., Rus, D., & Laubacher, R. (2020, December). *Artificial Intelligence and the Future of Work.* Retrieved from MIT Work of the Future: https://workofthefuture.mit.edu/wp-content/uploads/2020/12/2020-Research-Brief-Malone-Rus-Laubacher2.pdf

Matsumura, E., Prakash, R., & Vera-Munoz, S. (2014). Firm-Value Effects of Carbon Emissions and Carbon Disclosures. *The Accounting Review Vol. 89, No. 2.*, 695-724.

Melas, D., Nagy, Z., & Kulkarni, P. (2016, November). *Factor Investing and ESG Integration.* Retrieved from MSCI Research Insight: 2016 MSCI_Factor_Investing_and_ESG_Integration.pdf

Muguerza, N. M. (2014, 09). *El modelo de tres factores de Fama y French aplicado al mercado español .* Retrieved from Repositorio Universidad Pontificia Comillas: https://repositorio.comillas.edu/rest/bitstreams/2763/retrieve

Naciones Unidas. (2015). *Acuerdo de París.* Retrieved from UNFCCC: https://unfccc.int/sites/default/files/spanish_paris_agreement.pdf

Nielson, D., Nielsen, F., & Barnes, B. (2016). *An Overview of Factor Investing The merits of factors as potential building blocks for portfolio construction.* Retrieved from Fidelity Investments: https://www.fidelity.com/bin-public/060_www_fidelity_com/documents/fidelity/fidelity-overview-of-factor-investing.pdf

Open Bank & Santander Asset Management. (2021, 02). *Factor Investing, para una óptima diversificación, minimización del riesgo y búsqueda de retornos para las carteras*. Retrieved from Open Bank: https://www.openbank.es/open-news/factor-investing-carteras-fondos-inversion/

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). "Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques.". *Expert Systems with Applications 42 (1)*, 259–68.

Rapach, D., & Zhou, G. (2019, August). *Time-Series and Cross-Sectional Stock Return Forecasting: New Machine Learning Methods.* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3428095

Roncalli, T., Le Guenedal, T., Lepetit, F., Roncalli, T., & Sekine, T. (2020, July). *Measuring and Managing Carbon Risk in.* Retrieved from Research Gate: https://www.researchgate.net/publication/343879551_Measuring_and_Managing_Carbon_Risk_in_Investment_Portfolios?enrichId=rgreq-3c0035dbc0a73ae0da109012ed2a81ff-XXX&enrichSource=Y292ZXJQYWdlOzM0Mzg3OTU1MTtBUzo5Mjg3MjI0ODM5NTM2NjVAMTU5ODQzNjEwMzgzNg%3D%3D&el

Ulrich, E. (2016, December). *Entendiendo las inversiones según criterios ESG .* Retrieved from S&P Global Dow Jones Index: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.spglobal.com/spdji/es/documents/education/practice%20essentials-understanding-esg-investing-spa.pdf

United Nations Development Program. (2021). *World's largest survey of public opinion on climate change: a majority of people call for wide-ranging action.* Retrieved from Undp.org: https://www.undp.org/press-releases/worlds-largest-survey-public-opinion-climate-change-majority-people-call-wide-ranging-action

United Nations. (n.d.). *Macrodatos para el desarrollo sostenible*. Retrieved from Un.org: https://www.un.org/es/global-issues/big-data-for-sustainable-

development#:~:text=El%20volumen%20de%20datos%20en,del%20314%25%20respecto%20a%202015.

Universidad Pontificia Comillas & CFA Institute. (2022, Junio). *Portfolio management - chapter 2.* Retrieved from Moodle Rooms: https://sifo.comillas.edu/course/view.php?id=34640#section-1

Universidad Pontificia Comillas. (2019). Chapter 0 - Introduction to Machine Learning. *Machine Learning I*.

Universidad Pontificia Comillas. (2019). Chapter 10 - Neural Networks. *Machine Learning I*.

Universidad Pontificia Comillas. (2019). Chapter 11 - Support Vector Machines . *Machine Learning I*.

Universidad Pontificia Comillas. (2019). Chapter 8 - Decision Trees. *Machine Learning I*.

World Economic Forum. (2022, June). *Future Focus 2025 Pathways for Progress from the Network of Global Future Councils 2020–2022.* Retrieved from World Economic Forum: https://www3.weforum.org/docs/WEF_Future_Focus_2025.pdf