TÉCNICO LISBOA

Portfolio Selection and Trading Model based on Genetic Algorithms, K-Means Clustering, Fundamental Indicators and Technical Indicators

João Quinas Guterres

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor: Prof. Nuno Cavaco Gomes Horta

Examination Committee

Chairperson: Prof. Pedro Filipe Zeferino Aidos Tomás
Supervisor: Prof. Nuno Cavaco Gomes Horta
Member of the Committee: Dr. Nuno Calado Correia Lourenço

November 2022

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

First and foremost I would like to thank my supervisor, Prof. Nuno Horta, for giving me this opportunity, for always being available to discuss questions, suggest ideas and overall assist me with whatever problem I might have been faced with regarding the development of this work. I would also like to thank my colleagues that were developing their works simultaneously for always being open to share experiences and discuss ideas. Finally, I would like to thank my family and friends for always supporting me throughout this thesis, pushing me to continue working even when motivation might have been lacking.

Abstract

This thesis proposes a model capable of utilizing Evolutionary Algorithms and Fundamental Analysis for stock selection, this alongside a trading system based on Technical Analysis. Firstly a data scrapping program is created in order to obtain the financial statements of the S&P500 companies for every end of the month in the data range, allowing the calculation of various Fundamental Indicators. A K-means clustering algorithm is then used to classify these indicators among sector peers for every end of the month in the data range. A Genetic Algorithm will then search for the importance that should be given to each indicator using a previous three month stock price variation method as the fitness function. Three different strategies with different fundamental indicators are considered, one focused on growth, another on general performance and another focused on value companies. This stock selection algorithm is then tested under different portfolio compositions and alongside a trading system based on the EMA indicator, with the testing being done over the period of 2017-02-01 up to 2021-10-31. The results show that all strategies have potential, although the value and balanced strategies are the ones that most consistently beat the benchmark, yielding returns up to 119.4% in 2020, displaying their strength when recovering from market downtrends. The EMA trading system proved to be a useful tool as well, mainly when it comes to avoiding heavy losses and reducing volatility.

Keywords

Genetic Algorithms; K-means Clustering; Stock Selection; Trading System; Fundamental Analysis; Technical Analysis.

Resumo

Esta tese propõe um modelo capaz de utilizar algoritmos evolutivos e análise fundamental para seleção de ativos, acompanhado por um sistema de trocas de ações baseado em análise técnica. Primeiro um programa para obtenção de dados é criado de maneira a obter informação financeira relativa às empresas do S&P500 para cada fim do mês do intervalo de tempo considerado, permitindo o cálculo de diversos indicadores fundamentais. Um algoritmo de agrupamento K-means é utilizado para classificar os indicadores obtidos entre sectores para todos os fins de mês do intervalo de tempo considerado. Um algoritmo genético irá analisar a importância de cada indicator, sendo a função de aptidão um método baseado nas variações de preço de um ativo nos últimos três meses. Três estratégias diferentes foram testadas, cada uma considerando indicatores diferentes, uma focada em crescimento, outra em performance geral e a última em empresas mais seguras. Este algoritmo para seleção de ativos foi testado considerando vários portfólios diferentes e aplicando um sistema de trocas baseado no indicador EMA ao longo do período de 01-02-2017 até 31-10-2021. Os resultados mostram que todas as estratégias têm potencial, especialmente as focadas em equilíbro e empresas mais seguras, tendo em conta que são as que mais consistentemente batem o *benchmark*, chegando a ter retornos de até 119.4% em 2020, mostrando a capacidade de recuperar após quedas de mercado. O sistema de troca de ativos EMA provou ser uma ferramenta útil, maioritariamente no que toca a minimizar perdas e reduzir volatilidade.

Palavras Chave

Algoritmos Genéticos; Agrupamento K-means; Seleção de ativos; Análise Fundamental; Análise Técnica.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

2SP 2 Stock Portfolio

5SP 5 Stock Portfolio

7SP 7 Stock Portfolio

BBP Benchmark Beating Percentage

CAPM Capital Asset Pricing Model

CR Current Ratio

D/E Debt-to-Equity Ratio

DR Debt Ratio

EAs Evolutionary Algorithms

EMA Exponential Moving Average

EMH Efficient Market Hypothesis

EPS Earnings Per Share

FA Fundamental Analysis

FIs Fundamental Indicators

GA Genetic Algorithm

GICS Global Industry Classification Standard

GPM Gross Profit Margin

KOSPI200 Korea Composite Stock Price Index 200

MC Market Capitalization

MOEA Multi-Objective Evolutionary Algorithm

NI Net Income

NPM Net Profit Margin

NSGA-II Nondominated Sorting Genetic Algorithm II

NYSE New York Stock Exchange

OM Operating Margin

PER Price Earnings Ratio

PMP Positive Month Percentage

PR Payout Ratio

RG Revenue Growth

ROA Return On Assets

ROE Return On Equity

ROI Return On Investment

RR Retention Ratio

RSS Residual Sum of Squares

S&P 500 Standard and Poor's 500

SMA Simple Moving Average

SR Sharpe Ratio

SSE Shanghai Stock Exchange

SSP Sector Stock Portfolio

SVR Support Vector Regression

TA Technical Analysis

TIs Technical Indicators

WMA Weighted Moving Average

1

Introduction

Contents

This chapter will provide an introduction to the most well known investment strategies for stocks, how companies can be classified into different sectors and evolutionary algorithms. The motivation and objectives of the thesis will also be outlined alongside the the document's structure.

1.1 Initial Overview

1.1.1 Investment Strategies

The stock market encapsulates various venues where shares of publicly held companies can be bought, sold and issued, such as the New York Stock Exchange (NYSE), Shanghai Stock Exchange (SSE) and Euronext.

Investors purchase shares of a company in an attempt to profit via its stock price appreciation and regular dividend payments with the main goal being to maximize profits whilst reducing risk. Along the years multiple theories and strategies have been developed in order to thrive in this very volatile and competitive environment. Two of the most prominent schools of thought that have emerged are Fundamental Analysis and Technical Analysis.

Utilizing Fundamental Analysis, an investor seeks stocks that are undervalued expecting that their market price will eventually rise to its ”fair value”, capitalizing on the aforementioned opportunity. To determine the intrinsic value of a given stock numerous factors are considered, both qualitative, such as the management and business model, and quantitative, namely a company's financial statements and macroeconomic factors. On the other hand, Technical Analysis focuses on finding trading opportunities resorting to price trends, searching for patterns in charts and considering volume in an attempt to forecast price movements.

A popular theory that disregards the utility of both Fundamental Analysis (FA) and Technical Analysis (TA) is the Efficient Market Hypothesis (EMH) [1], which states that it is impossible to ”beat the market”, *viz.* stocks are already being traded at their ”fair value”, making it impossible to benefit from undervalued or overvalued stocks. Proponents of this theory believe that investing in a passive and low-cost market tracking portfolio is the best strategy one can find to profit in the stock market.

1.1.2 Sector Classification

When attempting to analyse the stock market it is important to aggregate companies with similarities into various groups, given that this allows the investor to get a better understanding of how a more specific type of company is performing, which may highlight general underlying problems or advantages. Various industry classification schemes have been developed, such as the Global Industry Classification Standard (GICS), the North American Industry Classification System (NAICS) and the Fama-French

(FF) industry classification, with the GICS being considered superior [2]. Developed in 1992 by MSCI and S&P 500 Dow Jones Indices, this classification scheme is a four-tiered system with the following hierarchy: 11 sectors, 24 industry groups, 69 industries and 158 sub-industries. Sectors defined by this system are: Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Communication Services, Utilities and Real Estate. Figure 1.1 displays a plot of sectors' performances alongside the Standard and Poor's 500 (S&P 500).

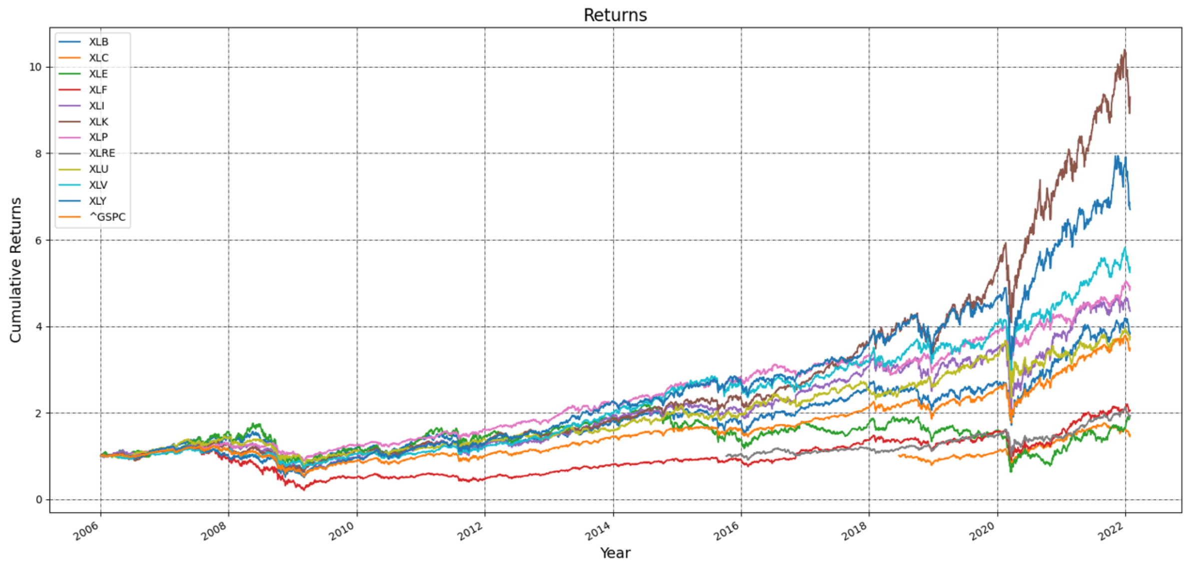

Figure 1.1: Plot of sectors' performances alongside the S&P 500 index. Labels: XLB - Materials; XLC - Communication Services; XLE - Energy; XLF - Financials; XLI - Industrials; XLK - Information Technology; XLP - Consumer Staples; XLRE - Real Estate; XLU - Utilities; XLV - Health Care; XLY - Consumer Discretionary; GSPC - S&P 500

1.1.3 Evolutionary Algorithms

In recent years computational intelligence has become more prevalent in the world of investing [3] [4] [5] , mostly due to the amount of information available regarding companies and financial markets. Many studies have been developed "covering techniques for preprocessing and clustering of financial data, for forecasting future market movements, for mining financial text information, among others", [3]. Technical and Fundamental indicators are usually considered, with the former being more commonly used as input for machine learning and deep learning models [4], although some studies advocate for the advantages of combining both indicators [6]. In this thesis both Technical and Fundamental indicators will be considered.

A particular type of algorithm that has gained traction is the Evolutionary Algorithm. Evolutionary Algorithms (EAs) [7] are metaheuristic optimization algorithms inspired by concepts in Darwinian Evolution with mechanisms that might include selection, reproduction, mutation and recombination. The

EAs family is composed by the following main algorithms: *genetic algorithm* (GA), *genetic programming* (GP), *differential evolution* (DE), the *evolution strategy* (ES) and *evolutionary programming* (EP).

The Genetic Algorithm (GA) is one of the oldest and most popular Evolutionary Algorithms. Based on nature, the GA takes into consideration the Darwinian theory of species evolution when searching for a solution space. In a population each individual, named a chromosome, will present itself as a possible solution for the problem at hand, with this problem being defined by the objective function. Each individual is attributed a value representing its quality depending on how well it is fitted to the objective function, referred to as the fitness of the individual, which plays the main role in evaluation. Individuals that carry higher quality are more probable to be selected to the new generation of a given population. Three operators are presented in a GA: selection, new population created based on fitness values; crossover, exchange of parts of individuals between two individuals; and mutation, specific genes are changed at random.

1.2 Motivation

As referenced before EAs have gained an important position in the world of investing with the main focus being on the application of these methods alongside Technical Indicators (TIs). The main motivation of this thesis is to look at the application of Evolutionary Algorithms through a different lens, testing a new application of GAs in investing, shifting the attention towards Fundamental Indicators (FIs). The objective isn't to delve into the novelties of Evolutionary Algorithms per se but rather search for a different method of utilizing these algorithms to the advantage of the investor.

1.3 Objectives

In this thesis, the main objective is to the develop a trading mechanism capable of selecting the best performing stocks and creating portfolios that can, at a bare minimum, outperform the S&P 500 benchmark.

In order to achieve this goal various secondary objectives must be met. Firstly it is important to obtain and treat fundamental data from S&P 500 companies in order to calculate various fundamental indicators. These indicators are then classified by contrasting them with sector peers. Utilizing a GA, an attempt is made to optimize the weights that each fundamental indicator must be given in order to find the best performing stocks within each sector of the GICS. Combining both the classified indicators and their respective weights companies can be ranked. Various portfolio distributions considering these novel rankings are then analyzed alongside a BUY/SELL trigger, all done aiming to maximize returns. The secondary objectives can be stated as:

- Download and use fundamental data in order to calculate various fundamental indicators for S&P 500 companies.
- Create a classifier that ranks a company by classifying its fundamental indicators, this based on contrasting these ratios against sector peers.
- With the help of a GA, determine weights that indicate the importance of each indicator.
- Combining both the classifier and weights create a novel stock ranking system.
- Take advantage of the new ranking system to select stocks for portfolio creation.
- Test these portfolios not only alone but also when applying a BUY/SELL system based on Technical Indicators.

1.4 Document Structure

This thesis is is organized as follows: Chapter 1 serves as an introduction to the topic approached, also addressing the motivation and objectives of this thesis. In chapter 2 a theoretical description of the concepts utilized during this thesis is provided alongside an overview of related works. In chapter 3 the overall architecture of the model generated is described, with an in depth explanation of all its components. Chapter 4 will provide a thorough analysis of various experiments made to better grasp the performance of the model described in chapter 3. Chapter 5 will go over the general conclusions, pointing out accomplishments and limitations and also suggesting possible improvements.

2

Background and Related Works

Contents

The Background chapter will go over the theoretical concepts utilized during this thesis. It will firstly focus on the two main investment schools of thought, Fundamental Analysis and Technical Analysis and then explain the main ideas behind Data Clustering and Genetic Algorithms. An analysis of related works will also be done, which will go over works on portfolio selection/optimization that utilize methods referred to in this thesis.

2.1 Fundamental Analysis

FA is, as mentioned above, a method of evaluating the ”fair value” of a given stock, considered to be initially adopted by Benjamin Graham [8] [9] and popularized by famed investor Warren Buffett [10]. The FIs used when analysing a company's fundamentals are usually of two types: *quantitative* or *qualitative*. When considering qualitative factors one tends to focus on the company's business model, *viz.* how is the company profiting exactly; on their management, analyzing the leaders' careers; on the corporation's industry, focusing on market share, regulations, industry size and competitors; and on the company's competitive advantage, e.g., powerful branding or monopolies, among other indicators. On the other hand, quantitative factors mostly rely on companies' financial statements, deriving many important indicators from the balance sheet, income statement and statement of cash flows. Although qualitative factors are of the utmost importance when an investor is considering to invest in a company, it is difficult to take advantage of them when constructing a trading model that resorts to optimization algorithms, thus they will not be considered in this thesis.

Quantitative FIs, and qualitative FIs for that matter, may also be divided in three different categories:

- **Macroeconomic**: Indicators that portray the current economic situation of a given country, region or sector [11]. Common indicators are the unemployment rate, the Gross Domestic Product (GDP) and the Consumer Price Index (CPI).
- **Industry**: Indicators that contrast the performance of a company among industry peers, and also analyse the current status of each industry.
- **Company**: Referring to the paragraph above, indicators derived from a company's financial statements.

Some commonly used factors obtained via financial statement analysis are:

- **Earnings Per Share (EPS)**: The quantity of a company's profit allocated to each share. A higher EPS indicates that the particular stock has greater value. An example of the evolution of the EPS alongside the share price of the ticker **MNST** can be found in figure 2.1.

$$EPS = \frac{NetIncome - DividendsonPreferredStocks}{AverageOutstandingCommonShares} \quad (2.1)$$

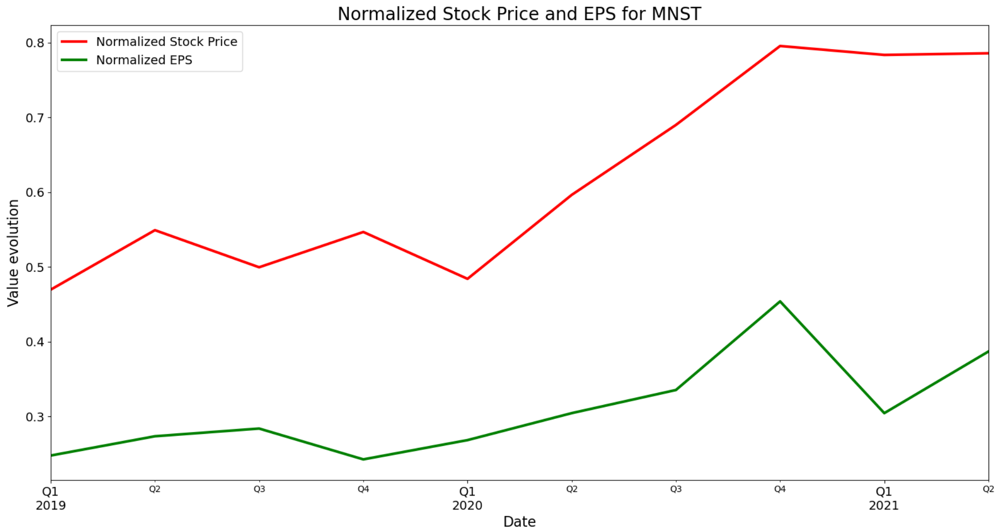

Figure 2.1: Normalized EPS and Share Price for the **MNST** ticker

- **Current Ratio (CR)**: Division between a company's current assets and its current liabilities.

$$\text{CR} = \frac{CurrentAssets}{CurrentLiabilities} \quad (2.2)$$

- **Debt Ratio (DR)**: Ratio that measures the level of debt of a company, portraying the amount of assets that are financed by debt.

$$DR = \frac{TotalDebt}{TotalAssets} \quad (2.3)$$

  If the ratio is high the company may be at risk of defaulting, while a ratio below 1 means that the company has more assets than debt. This ratio varies across industries, depending on how much capital is required.

- **Debt-to-Equity Ratio (D/E)**: Ratio calculated by dividing liabilities by shareholder's equity.

$$Debt/Equity = \frac{TotalLiabilities}{TotalShareholders'Equity} \quad (2.4)$$

  A higher value of this ratio is associated with a higher risk to the investor and it should be analysed at an industry level. The D/E is a good indicator of the type of strategy the company is adopting, where a higher value reveals that growth is being financed with debt, which can be associated with higher risk. An example of the evolution of the D/E alongside the share price of the ticker **AMZN** can be found in figure 2.2.

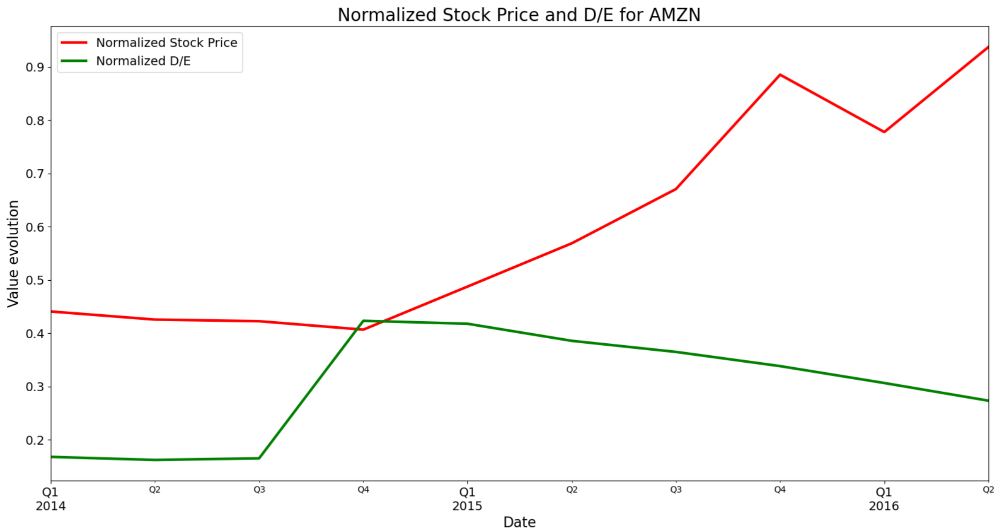

Figure 2.2: Normalized D/E and Share Price for the **AMZN** ticker

- **Return On Equity (ROE)**: Measures a company's Net Income (NI) divided by the company's total equity.

$$ROE = \frac{\text{NI}}{TotalEquity} \tag{2.5}$$

The ROE shows how profitable a company is, and, the higher the ratio, the higher the return made of the money invested on the stock. A company's ROE should be contrasted with its sector average, although a comparison with the general market is also valid.

- **Return On Assets (ROA)**: Divides the company's NI by its total assets.

$$\text{ROA} = \frac{\text{NI}}{TotalAssets} \tag{2.6}$$

This ratio allows the investor to get a better understanding of how effectively a company uses its assets when generation profit and also takes debt into consideration unlike the ROE. Its best to be compared between industries since some industries require more assets than others.

- **Retention Ratio (RR)**: Portion of returns that is kept to grow the business.

$$\text{RR} = \frac{\text{NI} - DividendsDistributed}{\text{NI}} \tag{2.7}$$

A higher RR usually indicates that the company under analysis is considered a growth company that has increased revenues and profit, although this ratio does not necessarily indicate that the company is reinvesting these funds.

- **Gross Profit Margin (GPM)**: This ratio demonstrates the profit made when only considering the cost of goods sold (COGS).

$$\text{GPM} = \frac{NetSales - COGS}{NetSales} \tag{2.8}$$

The GPM may be a useful indicator when trying to determine how much a company is profiting against the costs of production, which generally reflects on good management.

- **Net Profit Margin (NPM)**: This ratio measures the margin (profit) the company has after paying all the operating, administrative and financial costs, along with taxes.

$$\text{NPM} = \frac{\text{NI}}{Revenue} \tag{2.9}$$

The NPM allows the investor to assess the capacity of a company to profit, being one of the best ratios for detecting potential. If the profit margin is zero or negative one can conclude that the business is either not managing their expenses or not selling enough goods. An example of the evolution of the NPM alongside the share price of the ticker **TSLA** can be found in figure 2.3.

- **Revenue Growth (RG)**: Shows the evolution of a company.

$$\text{RG} = \frac{Revenue_{Current} - Revenue_{LastYear}}{Revenue_{LastYear}} \tag{2.10}$$

If the revenue of a company is increasing it is either inserted in a growing sector or getting the better of competition. An example of the evolution of the RG alongside the share price of the ticker **MRNA** can be found in figure 2.4.

- **Payout Ratio (PR)**: Indicates the percentage of net income distributed by the investor as dividends.

$$\text{PR} = \frac{DPS}{\text{EPS}} \tag{2.11}$$

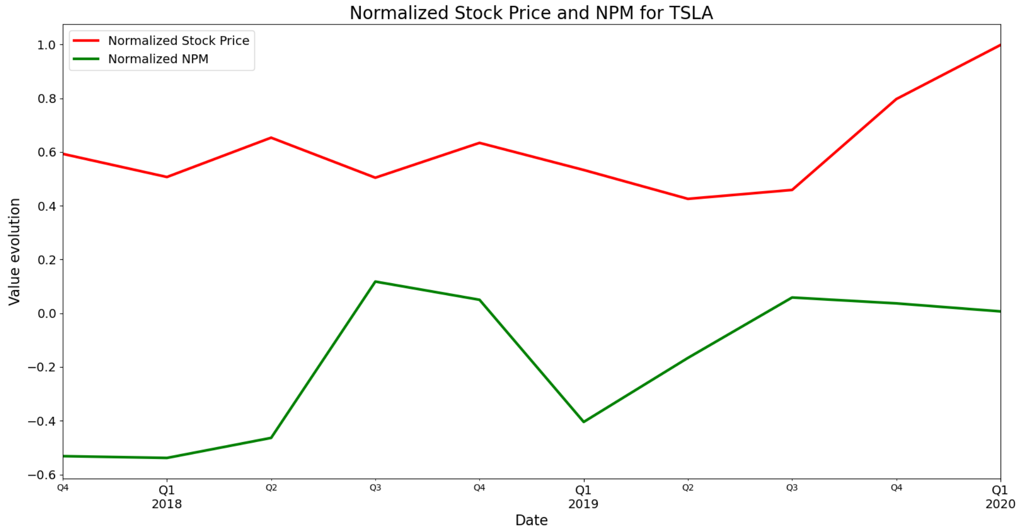

Figure 2.3: Normalized NPM and Share Price for the **TSLA** ticker

If an investor is looking to invest in a growth focused company, he should seek a lower PR ratio, since a lower value reflects that the company is reinvesting its excess capital rather than distributing it among shareholders. Yet, if seeking for higher income, the investor should search for a higher PR ratio.

- **Operating Margin (OM)**: Indicator that relates the profit made via operations with total revenues, which allows the investor to grasp the main source of profit for the company.

$$\text{OM} = \frac{OperatingEarnings}{Revenue} \tag{2.12}$$

This ratio also allows the investor to understand the proportion of revenue that can be used to pay expenses such as interest. Higher values are considered positive and a high variance of this ratio suggests higher risk.

- **Price Earnings Ratio (PER)**: Ratio that indicates the value of a company's share price when compared with its EPS.

$$\text{PER} = \frac{SharePrice}{\text{EPS}} \tag{2.13}$$

A higher PER may sprout from one of two situations, either the company is overvalued or is expected to grow, whilst a lower PER indicates that the company is either undervalued or performing exceedingly well. It is advisable to compare PERs between same sector companies.

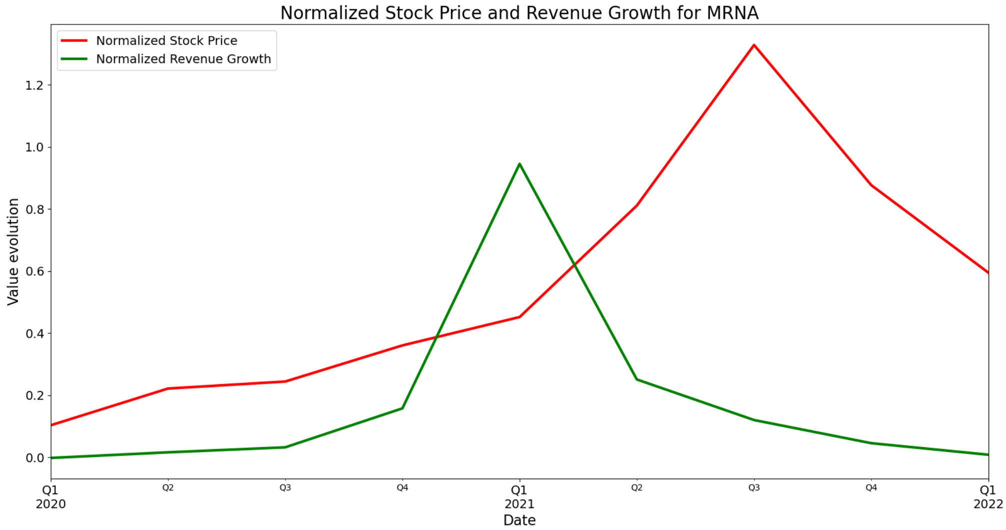

Figure 2.4: Normalized RG and Share Price for the **MRNA** ticker

- **Market Capitalization (MC)**: The total market value of all the company's shares.

$$\text{MC} = Company'sShares \times MarketPrice \quad (2.14)$$

Important metric that allows the investor to understand the impact of a given company on its market. If the investor is looking for smaller companies, usually more volatile but with more growth potential, he seeks for a smaller MC, whilst if searching for more stable companies a bigger MC is sought.

2.2 Technical Analysis

As stated previously, TA [12] takes advantage of the statistical analysis of trends when searching for trading opportunities. TIs, used by the technical analyst when predicting future price movements, are heuristic or pattern-based signals based on price and/or volume. Following [5], TA can be divided into eight main groups: Sentiment, Flow-of-funds, Raw data, Trend, Momentum, Volume, Cycle and Volatility. In this thesis the group highlighted will be **Trend**, which is price-based, searching for stock trends caused by investors' reactions to particular global events. The most well-known trend indicators are Moving Averages such as the three described bellow:

- **Simple Moving Average (SMA)**: The SMA is a simple indicator that aims at determining future price trends by averaging the previous closing prices of a security over a predetermined period. The formula is as follows:

$$SMA = \frac{S_1 + S_2 + ... + S_n}{N}, \tag{2.15}$$

where S_n is price at the period n and N is the number of periods. The SMA provides a better view of a security's price trend since it levels volatility.

- **Weighted Moving Average (WMA)**: This indicator gives more value to recent data points, reducing their significance as the distance increases. The formula for the WMA is:

$$WMA = \frac{S_1 * N + S_2 * (N-1) + ...S_n}{\frac{N*(N+1)}{2}} \tag{2.16}$$

where N is the time period and S_n is the price at period n

- **Exponential Moving Average (EMA)**: The EMA is a moving average that has an increased focus on recent days, thus being more sensitive to recent price changes. The formula for this indicator is as follows:

$$EMA_t = (S_t * (\frac{s}{1+D})) + EMA_y * (1 - \frac{s}{1+D}) \tag{2.17}$$

where EMA_t is the EMA today, EMA_y is the EMA yesterday, S_t is the price today, D is the number of days considered and s corresponds to the smoothing factor, which gives the recent value more weight, being typically 2. As this value is increased so is the influence of the most recent observations.

2.3 Data Clustering

The main objective of data clustering [13] is to naturally group sets of patterns, points or objects. In general clustering can be defined as: Given n objects, search for K clusters based on similarity, where the objects have high similarities inside the same clusters and low similarities with objects outside their clusters. The unknown amount of possible clusters and large data sets, both in quantity and dimension, require the usage of algorithms to obtain a better grouping of objects. Data clustering is present in various fields and finance is no exception [14] [15] [16] [17]. Following [13], clustering algorithms can be either **hierarchical** or **partitional**.

- **Hierarchical**: These algorithms find nested clusters by either initially making each point its own cluster and then consecutively joining the most similar pair of clusters, or considering that all initial point belong to one cluster and then separating clusters into smaller ones.

- **Partitional**: In these cases all clusters are found at the same time via data partition, with no hierarchical structure assumed. K-means clustering belongs to this group.

2.3.1 K-Means Clustering

The K-Means clustering algorithm [18] is one of the most basic and efficient data clustering algorithms. Initially the algorithm must select K points as the initial centroids of clusters. Based on a proximity metric each point is given its closest centroid and, after the clusters are formed, the centroids for each cluster are updated. These two steps will then be repeated until the centroids for each cluster are no longer updated. A simple example of the outcome of the K-means Clustering algorithms can be found in figure 2.5.

Algorithm 2.1: K-Means Clustering

1: Select K points as initial centroids
2: **While** convertion criterion is not met:
3: Form K clusters by assigning each point to its closest centroid
4: Recompute the centroid of each cluster

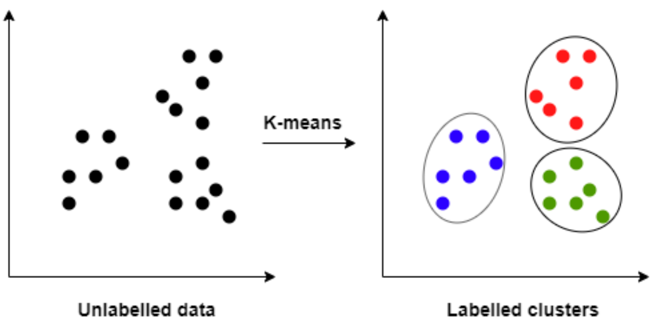

Figure 2.5: K-means outcome example

Various different proximity metrics may be used, such as the Manhattan distance, the Euclidean distance and Cosine similarity, with the Euclidean distance being the most common one. The objective function used by K-means is the Residual Sum of Squares (RSS):

$$RSS(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - c_k||^2 \tag{2.18}$$

where, for a dataset $D = \{x_1, x_2, ..., x_N\}$ with N points, the clustering obtained after applying K-means is $C = \{C_1, C_2, ..., C_k, ..., C_K\}$. c_k corresponds to the centroid of cluster C_k and can be defined as:

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|} \tag{2.19}$$

The objective is find a clustering that minimizes the RSS.

2.4 Genetic Algorithms

Before analysing how a GA [19] operates it is important to clarify some terms that are used in Evolutionary Algorithms.

- **Population**: Set of *individuals* of a given size.
- **Individuals**: Sets of parameters coded in the form of chromosomes corresponding to solutions, also known as the *search space points*.
- **Chromosomes**: Ordered sequences of *genes*.
- **Gene**: Consists of a single element of a chromosome in particular.

In GAs the *fitness function*, also referred to as the *objective function*, will measure the fitness of each individual in a population, and, following evolutionary principles, evaluates the fitness of all individuals, selecting the ones that are best fit. A basic GA, summarized in figure 2.8, encompasses the following steps:

- **Initial Population**: Random selection of all the individuals in a population, creating the initial population.
- **Evaluation**: Calculating the fitness of each chromosome in the population. The fitness function depends on the type of problem but it is expected to always take nonnegative values.
- **Stopping Criterion**: The stopping criterion is determined depending on the application of the algorithm. If the optimal value is known the algorithm may only be stopped upon its achievement, possibly with a predetermined accuracy. The algorithm may also be stopped when no further improvements are found. In other cases a maximum number of generations or a defined period of time may be set as the stopping criterion. If the algorithm stops, *viz.* the stopping criteria is met, the best chromosome is presented, else we enter chromosome selection.
- **Selection**: Based on their fitness, the chromosomes that participate in the creation of the new generation will be selected, following the rules of natural selection where the fittest individuals have a higher chance of participating in the generation of new chromosomes. The most common selection method is the roulette wheel selection method, that begins by assigning to each chromosome a sector of the roulette wheel proportional to the value of its fitness function. Each chromosome

denoted by ch_i for $i = 1, 2, ..., K$, where K signals the size of the population, corresponds to a given wheel sector $v(ch_i)$ of the entire wheel, represented by the following formula:

$$v(ch_i) = p_s(ch_i) \times 100\% \tag{2.20}$$

where

$$p_s(ch_i) = \frac{F(ch_i)}{\sum_{j=1}^{K} F(ch_j)} \tag{2.21}$$

with $F(ch_i)$ representing the value of the fitness function of chromosome ch_i and $p_s(ch_i)$ representing the probability of selecting the corresponding chromosome. Naturally, the larger the sector, the higher the probability of selecting the corresponding chromosome.

The selection process generates the so called *parents population* with the size of the population.

- **Genetic Operators**: After the parents population is selected, the application of genetic operators will create the new generation. The basic generic operators are the *crossover operator* and the *mutation operator* with the latter playing a secondary role. The mutation may occur before the crossover operation, *viz.* on the parents population, or after the crossover operation, on the resulting offspring.

- **Crossover**: Firstly the chromosomes from the parents population are mated in pairs at random, following the probability of crossover p_k. For each pair the gene position in the chromosome is chosen, defining the *crossover point*. Given that the chromosome of each of the parents is of size L then the crossover point $l_k \in [1, L-1]$. The following pair of offspring is then created:

  – Offspring where the chromosome is made of genes from the first parent from 1 to l_k and, from $l_k + 1$ to L, from genes of the second parent.

  – Similarly, offspring where from 1 to l_k the genes from the second parent are used and, from $l_k + 1$ to L, genes from the first parent are selected.

For example, considering the two parents shown in figure 2.6 and $L = 4$ we obtain the pair of offspring represented in figure 2.6.

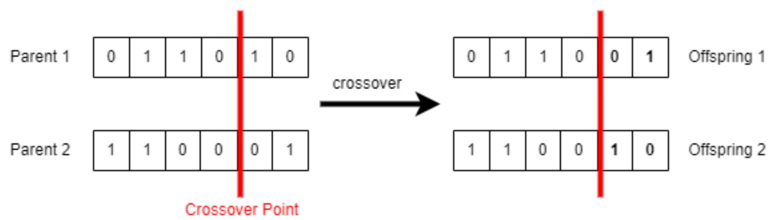

Figure 2.6: Crossover example

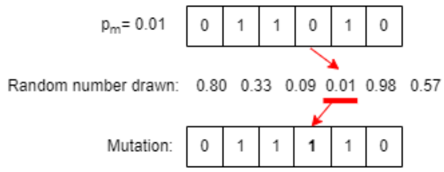

Figure 2.7: Mutation example

- **Mutation**: Considering the probability of mutation p_m, the mutation operator replaces the gene value in the chromosome to the opposite value. p_m is usually very small and, for a mutation to occur a number from the interval [0,1] is drawn for each gene and, if this number is less or equal to p_m, the given gene is mutated.

  For example, considering the chromosome shown in figure 2.7 and a $p_m = 0.01$ the represented mutation is obtained.

- **New population creation**: The new population is then obtained, resulting from the application of the genetic operators to the parents population. For each generation the fitness function's value for all chromosomes must be calculated. The stopping criterion is checked and, if met, the fittest chromosome is presented, else, the algorithm moves onto the next step, which is selection. The classical genetic algorithm replaces the previous population with a new population of offspring of equal size.

- **The best solution**: When the algorithm stops, the best solution, corresponding to the fittest chromosome, is presented.

2.5 Related works

When searching for related works the main focus was to select articles that had some correlation with the objectives of this thesis. Thus the focus of this search was towards the following topics: Evolutionary algorithms in finance, especially focused o on portfolio composition and paired with fundamental indicators; the usage of clustering for stock selection; and trading systems that take advantage technical indicators.

In [20] a GA was taken advantage of in order to optimize portfolios with the assistance of a momentum strategy and using the Capital Asset Pricing Model (CAPM) to determine undervalued stocks. The implemented fitness function took into account the Portfolio fund standardization, the Portfolio CAPM and its Sharpe Ratio when evaluating chromosomes. The model was tested in two different markets, S&P 500 and Korea Composite Stock Price Index 200 (KOSPI200), projecting a better performance than both indexes.

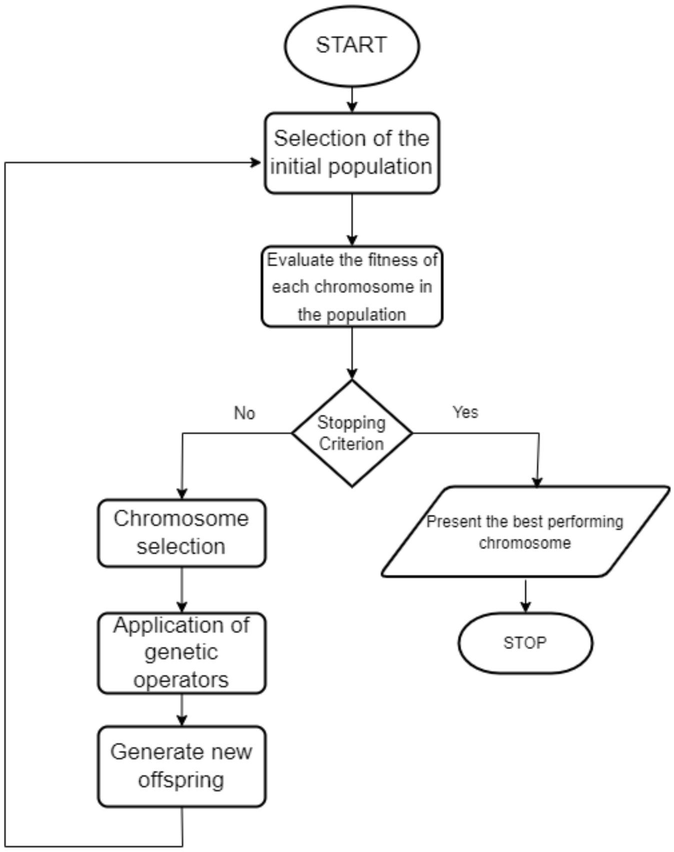

Figure 2.8: Flowchart of a generic Genetic Algorithm

In [21] both financial ratios and technical indicators were used, alongside a Multi-Objective Evolutionary Algorithm (MOEA) (SPEA II). The fitness functions considered were the return and its respective risk, and the models considered up to 10 financial ratios and 5 trading parameters for the chromosomes. Tested for the S&P 500 from June 2010 to 2014, the results obtained were not only above the market's average but also paired with low variances.

In [22] a clustering-based portfolio optimization model using a GA and investor information is implemented. Firstly, taking advantage of investor information, various portfolios are generated via clustering analysis, using K-means clustering. Then, in order to optimize the weights of the selected stocks, a GA is used, with the fitness functions being either the Minimum variance weights or the Sharpe ratio weights. The proposed model outperformed previous models when applied to the KOSPI200.

In [23] a MOEA (Nondominated Sorting Genetic Algorithm II (NSGA-II)) portfolio optimization model is paired with TIs. The genetic algorithm aims to minimize risk (Covariance or CVar) and maximize the return function, developed in [24], also considering 4 indicators. Two scenarios are possible, the first one only using the TIs for transactions after the optimal portfolio is chosen at the start of each month and the second will perform the monthly optimization only on the stocks selected by the indicators. The simulation encapsulates 6 years of data from the Brazilian Stock Exchange and the strategy focused on using the optimization first performed better than both the index and the other strategies.

In [25] a Support Vector Regression (SVR) method generates replacements for actual stock returns, in order to rank stocks, where the top rated stocks will be used when forming a portfolio. Supporting the SVR, a GA is employed for parameter optimization, and feature selection providing the best input variables to the SVR model. The data selected corresponds to the 200 largest market capitalization stocks in the Taiwan Stock Exchange, with dates ranging from 1996 to 2010. The attributes used in the stock selection model are fundamental ratios related to share price rationality, profitability, leverage, liquidity, efficiency and growth. The models applied significantly outperformed the benchmark, achieving a maximum 17.5719 % mean of annualized model return.

In [26] TIs and FIs are used for the creation of a model, alongside a MOEA that assesses risk and returns for optimization. The study was based on 40 stocks of the SSE A, obtained via the Cathay Pacific database. The results show that the model created is able to outperform the index and, as the number of fundamental indicators increases so does the performance, specially with the addition of cash-flow growth, capital expenditure growth rates and the payout ratio.

In [14] a portfolio construction method is proposed that takes into account the continuous trend characteristics of the market. Firstly K-means clustering is used to cluster stocks, divide the different stock groups and revise the calculation of returns for the Sharpe Ratio, this based on the continuous trend characteristics. Various portfolio theories are then utilized to calculate the required weights. The results obtained show that the proposed method was superior.

In [16] a portfolio selection algorithm is presented that, based on the pattern matching principle, selects the optimal portfolio, this updated periodically. The two steps of the system consist of the sample selection, that utilizes various clustering algorithms including k-means, and the portfolio optimization, where the optimum function and transaction costs are considered. Various data sets were considered for different american indexes and time frames, ranging from 2001 up to 2017. The results indicate that the proposed models outperformed others provided in the literature.

In [27] the performances of the SMA, EMA, MACD and Triple Screen techniques are analyzed in a trading system. 198 stocks traded in the Brazilian stock market were used for various different periods ranging from 2000 up to 2014. Multiple brokerage fees were considered alongside a Stop-Loss mechanism and the benchmark considered was a buy-and-hold strategy. Although most results indicated that the strategies had positive returns only a small percentage was actually capable of overcoming the buy-and-hold strategy.

Table 2.1: Related works

Work	Date	Algorithm	Fitness Function	Benchmark	Financial Application	Period	Results
[20]	2020	GA	Sharpe Ratio and Portfolio CAPM	S&P 500; KOSPI200	Portfolio Optimization	2008-2018	Best: 400% cumulative returns
[21]	2015	MOEA (SPEA II)	Returns and variance of returns	S&P 500	Portfolio Composition	2010-2014	Best: 50.24%
[22]	2017	GA	MV weights/Sharpe weights	KOSPI200	Portfolio Optimization	2007-2014	Best: Annual Return of 40.33%
[23]	2021	NSGA-II	Returns and Risk	Brazilian Stock Exchange	Portfolio Optimization	2012-2015	Best: 68.09%
[25]	2012	SVR and GA	Annualized Return of the Portfolio	Taiwan Stock Exchange	Stock Selection	1996-2010	Best: 17.5719 %
[26]	2018	SPEA-II	Returns and Risk	SSE	Portfolio Optimization	2012-2013	Best: 47.07%
[14]	2022	K-means	-	Chinese stock markets	Portfolio Construction and Optimization	2001-2020	Annualized Returns: 30%
[16]	2020	Various clustering algorithms	-	Multiple American Indexes	Stock Selection and Portfolio Optimization	2001-2017	Best: 45.09%
[27]	2015	-	-	Brazilian Stock Market	Trading System	2000-2014	Best average return (day): 0.061%

3

Architecture

Contents

This chapter will go over the architecture and implementation of the model created. After a global view every module utilized will be thoroughly described in order to properly understand the implementation.

3.1 Global View

The various modules of the architecture are defined in figure 3.1.

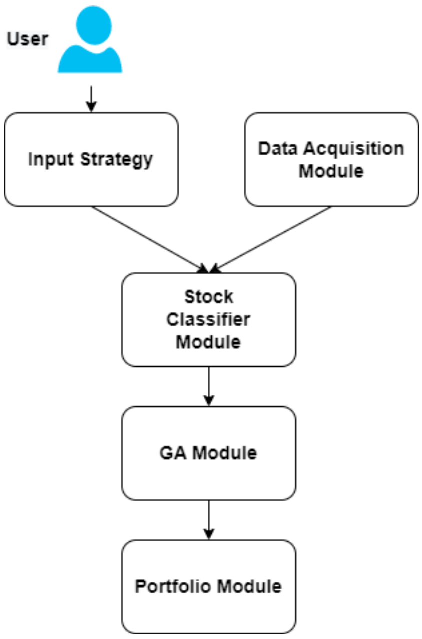

Figure 3.1: Global view of the architecture employed

A small summary of each module will be given.

- **Input Strategy**: The investor may choose between three main strategies, a **Growth Strategy**, a **Balanced Strategy** and a **Value Strategy**. The strategy chosen will influence the indicators taken into consideration when classifying stocks and training the GA.
- **Data Acquisition Module**: This module will be responsible for the storage of S&P 500's companies fundamental information as well as their historical price variations. Information relative to fundamental indicators and stock price variations is saved for the tickers considered (where each ticker represents a specific company), this for the entire data range.
- **Stock Classifier Module**: Based on the strategy chosen, this module will utilize the fundamental indicator information obtained in the **Data Acquisition Module** and classify them for all companies considered at the end of each month via a clustering method. For each ticker and end of the month these classifications will be saved.

- **GA Module**: Depending on strategy, this module will attempt to correlate the importance of each indicator with the historical price variations. In order to achieve this the GA will need the classifications obtained in the **Stock Classifier Module** and the stock price variations from the **Data Acquisition Module**, with the output being the best chromosome for each of the training periods considered, which will contain the weights given to each indicator.
- **Portfolio Module**: This module will generate various portfolios based on the best companies gathered by combining the weights obtained in the **GA Module** and the classifications derived from the **Stock Classifier Module**. Pairing the classifications and weights a new ranking system is used to select companies for portfolio creation. The usage of the technical indicator EMA will also be considered to decide IN/OUT positions.

Figure 3.2 demonstrates the general flow of information.

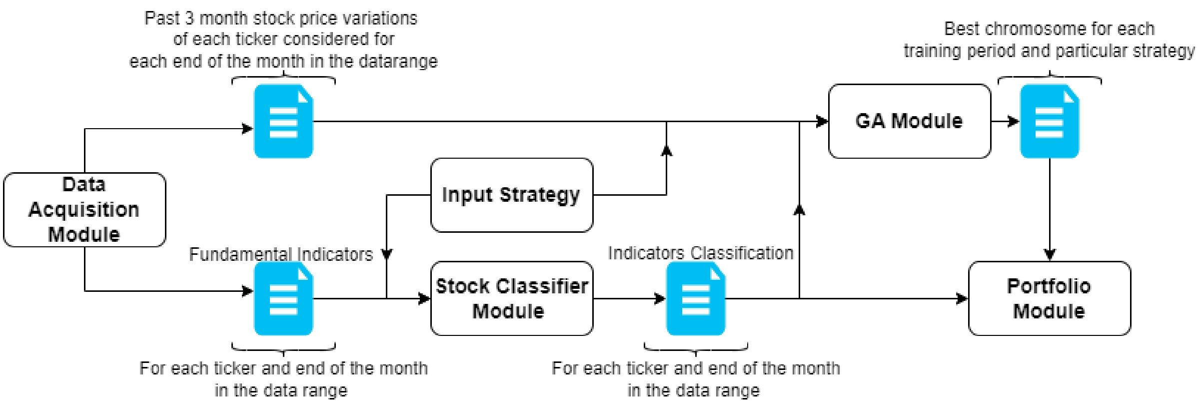

Figure 3.2: General flow of information

3.2 Input Strategy

The investor may choose various strategies that will define which indicators will be considered in the GA.

- **Strategy A - Growth**: The indicators taken into consideration will be more focused on growth, overlooking debt to some extent and searching for growing and undervalued stocks that appear to have a good structure/business model. A higher RR will be sought since it indicates that the returns are being reinvested in the company. The NPM ratio is also taken into consideration since it allows the investor to better understand the capacity of a company to profit, an indicator of potential. RG is naturally a considered indicator in this strategy given that it shows if revenues are increasing, a characteristic of growth companies. The D/E is calculated by dividing liabilities by shareholder's equity so, if this value is higher compared to industry peers it means that the company's growth is being financed by debt, also increasing risk. Lastly, the ROE is considered for this strategy since, unlike the ROA, it ignores debt, solely focusing on profits.

- **Strategy B - Balanced**: A balanced set of indicators is considered, taking into account both growth and stability. The D/E will be considered, as explained before an indicator focused on the financing of growth. The CR is obtained by dividing current assets by current liabilities, showing a company's capability of covering short-term debt, being stronger when compared at an industry level. One of the most well-known indicators, the EPS, is particularly useful when looking for the value in companies. Both the ROE and ROA will be considered in this strategy, the main difference being the consideration of debt. The GPM and NPM will also be indicators used by this particular strategy, with the former focusing on good management and the latter highlighting possible potential. RG will be taken into account as a growth indicator and the OM, that relates profit from operations with revenue, will also be considered. Overall the idea is to find a more balanced fundamental view of the S&P 500's companies.

- **Strategy C - Value**: The main objective is to search for companies that have a dominant position in their respective sector, presenting good debt management, sustained growth and solid foundations. When considering these qualities the EPS ratio is a great indicator of value. Another useful ratio is the CR that, as explained before, indicates the strength of a company when covering short-term debt and important trait when looking for a stable company. The ROA and RG are again considered, the former taking into account debt which is always important in value companies and the latter searching for companies that are still profitable/have strong positions in the market. Finally, the PR is considered, given that it indicates income given to investors.

3.3 Data Acquisition Module

The Data Acquisition Module, defined in figure 3.3, is mainly focused on two tasks: Gathering information regarding the fundamental indicators of each ticker for the data range considered and obtaining the monthly stock prices of tickers, calculating the past three month stock price variations and ranking the companies for each end of the month in the data range based on these variations.

Fundamental Indicators: In order to obtain the fundamental indicators of all S&P 500 companies the **Macrotrends** website [28] was taken advantage of, alongside Python. Firstly a Python program that obtains financial quarterly data (income statement, balance sheet, cash flow statement and key financial ratios) from each S&P 500 company was developed. Then each one of these tables was saved into a csv file, where the income statement goes to *TICKER-income.csv*, the cash flow statement to *TICKER-cf.csv*, the key financial ratios to *TICKER-fr.csv* and the balance sheet to *TICKER-balance.csv*. The

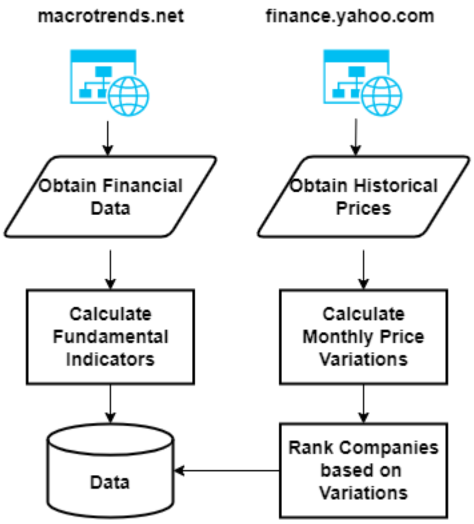

Figure 3.3: Overview of the Data Acquisition Module

data had the maximum range [31-03-2005, 30-09-2021], although some tickers do not have such a wide range. Tickers that were not found were discarded.

After saving all these statements, another Python program was created with the aim of calculating, for each company and quarter, various financial indicators and saving them into csv files of the type *TICKER-ratios.csv*. For each quarter, resorting to the yfinance library [29], the stock price of each ticker was also saved. If a stock price was not found for that specific date, due to it being on a weekend/holiday, the program goes back one day until a value is found. Stocks with varying data size in financial statements were discarded. The overall process is described in figure 3.4.

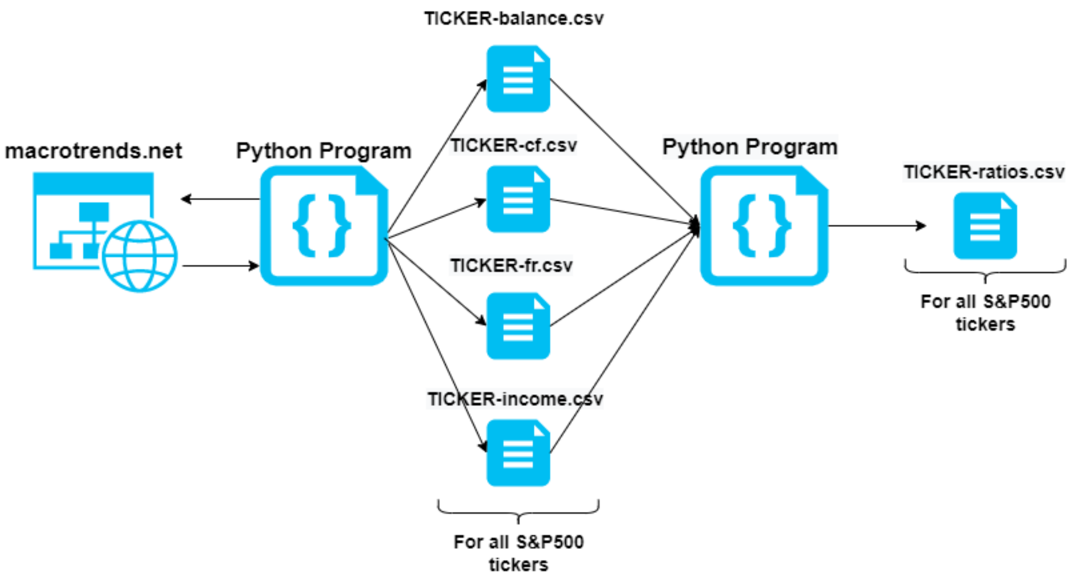

Figure 3.4: Schematic of how fundamental indicators are obtained

Algorithm 3.1: Data Scraper - Pseudo-code

```
Obtain all tickers used and corresponding names
function get_financial_data(tickers, names, number_tickers)
    for each ticker:
        Obtain the urls necessary to access the ticker's financial statements
        for each url:
            function request_data(current_url)
                return dataframe of statement information
            Save csv file regarding that ticker and financial information like shown in figure 3.4
```

Algorithm 3.2: Fundamental Ratio Calculator - Pseudo-code

```
Get all tickers considered
for each ticker:
    Open the income, balance, cf and fr csv files for that ticker
    Convert the files into dictionaries
    if the values are consistent:
        function get_fundamental_indicators(ticker, statement_dictionaries)
            Save the dates and calculate and save the fundamental indicators for those dates
            return dictionary with dates and calculated indicators
    else discard the ticker
    Save a csv file for that ticker as exemplified in figure 3.4
```

Three Month Price Variations: The gathering of historical prices was also done with the aid of the yfinance library [29]. A Python program went through each ticker and saved the stock value at the end of each month. If the date for the end of the month was not found then the program went back a day until the closest date with an associated price was found, saving that value. Utilizing these values the past three month variations of each ticker were calculated for the entire data range, permitting a monthly ranking of tickers based on the performance in the past three months.

3.4 Stock Classifier Module

The overall structure of the **Stock Classifier Module** is defined bellow in figure 3.5. The main objective of this module is to classify, for each date in the data range and for each ticker considered, the fundamental indicators of all companies included, with the FIs selected depending on the strategy chosen in section 3.2. To achieve this objective the tickers were distributed into sectors and their fundamental indicator values stored with sector peers for all dates and FIs. These values were then clustered utilizing a K-means clustering algorithm and classified from 1-20, providing, for each date and ticker, the various classifications of their FIs.

A Python program was developed that initially loads the indicators obtained in figure 3.4 for all tickers. Figure 3.6 displays an example of the structure of the files that are loaded. Considering that **N** fundamental indicators are calculated for each ticker in the **Data Acquisition Module** over **T** dates, the

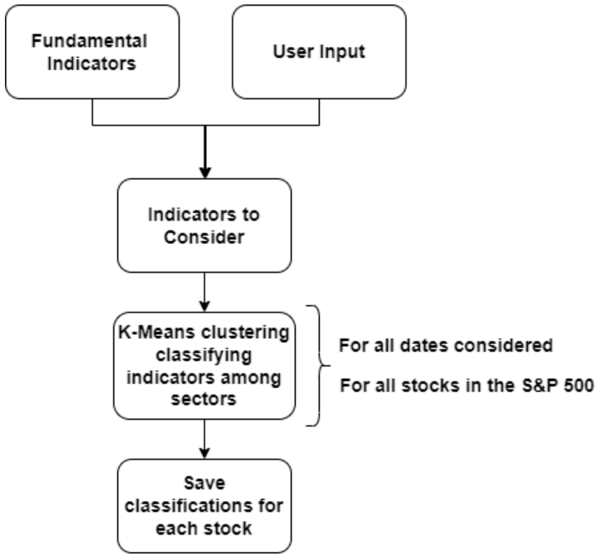

Figure 3.5: Overview of the Stock Classifier Module

structure will have for each date (column **date**) the current value of each fundamental indicator for a particular ticker (columns FI_n) which is defined as $Value_{tn}$ where t signals a row between 1 and T and n signals a column between 1 and N.

TICKER_ratios.csv

	date	FI_1	FI_2	FI_3	...	FIN
Row 1	2021-09-30	Value_11	Value_12	Value_13	...	Value_1N
...	...	...	...	...	...	...
...	...	...	...	...	...	...
Row T	2005-03-31	Value_T1	Value_T2	Value_T3	...	Value_TN

Figure 3.6: Structure of the *TICKER-ratios.csv* file

Now let it be considered that the **Strategy A**, defined in section 3.2 only considers the fundamental indicators FI_1, FI_2 and FI_3, **which is not true and merely exemplary**, then the program would only load the first four columns of the file described in figure 3.6, that is, columns **date**, FI_1, FI_2 and FI_3. After the correct indicators are loaded in accordance with the current strategy all the tickers must be grouped among their respective GICS sectors, with the illustration of this grouping being shown in figure 3.7. This is done because for all tickers, for each date, each indicator value will be compared among the values of sector peers. Firstly all tickers will be separated to 11 different pools depending on their sector, then among each pool the values of each fundamental indicator will be sorted among other tickers within the pool. Figure 3.8 represents the values of fundamental indicator FI_1 for the Materials Sector on a given end of the month **as an example**.

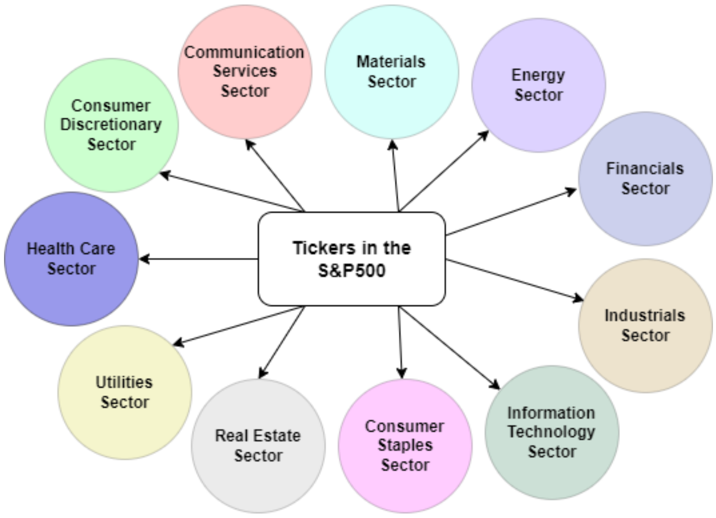

Figure 3.7: Distribution of stocks among the various GICS sectors

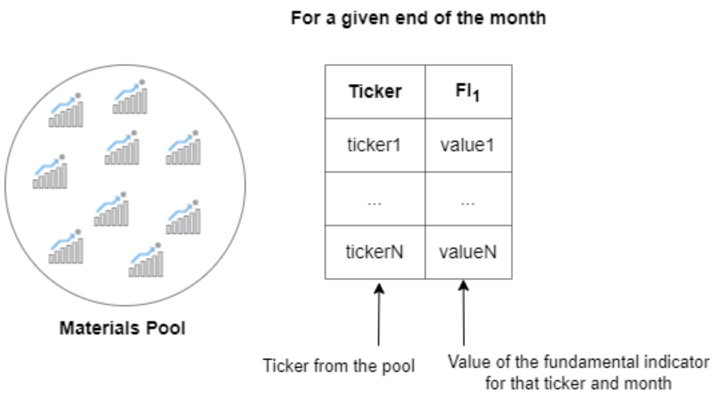

Figure 3.8: Values of FI_1 for the Materials Sector on a given month (Purely exemplary)

Continuing the **example** these values must then be sorted from best to worst in order among this sector. Naturally this process of sorting values must be done for all indicators being considered, for all tickers and for the entire data range.

Upon completion of the sorting, classification is ensued. The objective of this classification is to give a 1-20 score to each ratio for each company and date, where the performance of a given indicator is contrasted with sector peers and classifications of 0 represent lack of information. The method applied is K-means clustering with 20 clusters that takes advantage of the sorting explained. This way, the grouping of stocks in clusters is not only based on the value of the indicator but also on the overall distribution of values, with the higher values getting a better classification. It is again important to note that clusters are only created among companies that have the same sector

In some cases, the number of companies under analysis is bellow 20, thus an adaptation to the classification method is required. The number of clusters will then correspond to the number of stocks in the pool and instead of a classification dropping from 20 to 19 it will drop from 20 to $20 - \frac{20}{NumberofClusters}$.

Finally, a csv file of the type *TICKER-classkmeans-STRATEGY*.csv which saves the classification for each indicator for all available dates for the stock **TICKER** is generated for all tickers considered, following the format in figure 3.9. Each strategy will consider different indicators thus each ticker will have three different files. The data range taken into account is **31-03-2005** - **30-09-2021**. Overall, the classification of FIs can be simplified as in figure 3.10.

TICKER-classkmeans-STRATEGY.csv

Date	FI_1	...	...	...	FI_n
2021-09-30	13	...	...	...	20
...	...	...	...	...	...
...	...	...	...	...	...
2005-03-31	4	...	...	...	12

Figure 3.9: Structure of classification csv type (values are exemplary)

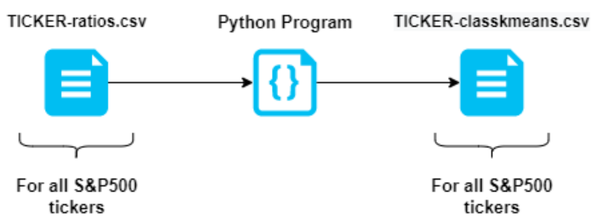

Figure 3.10: Input and Output of the classifier

The K-means algorithm used was from the scikit-learn library [30] and the pseudo-code describing the process of obtaining the *TICKER-classkmeans.csv* is described in 3.3.

Algorithm 3.3: Stock Classifier Module - Pseudo-code

```
Obtain all tickers that will be considered
Obtain the fundamental indicators calculated for each ticker in the Data Module
Register the dates and indicators for each ticker
Organize information by date, saving information regarding tickers, their sectors and
  corresponding indicator values for each date
for each date:
  function analyzing(information(date), companies_by_sector)
    for each indicator:
      function classification_per_ratio(information(date), indicator, companies_by_sector)
        Sort tickers' indicator values by performance
        Classify all tickers per sector using K-means
        return the classifications of all tickers for that date and indicator
    return the classifications obtained for all indicators for that particular date
  Save the dates and corresponding indicator classifications for each ticker
Create a csv file following the format presented in figure 3.9 for each ticker, saving the
  indicator classifications for each date
```

3.5 GA Module

The overall functioning of the GA Module is shown in image 3.11. A step by step explanation will now be done.

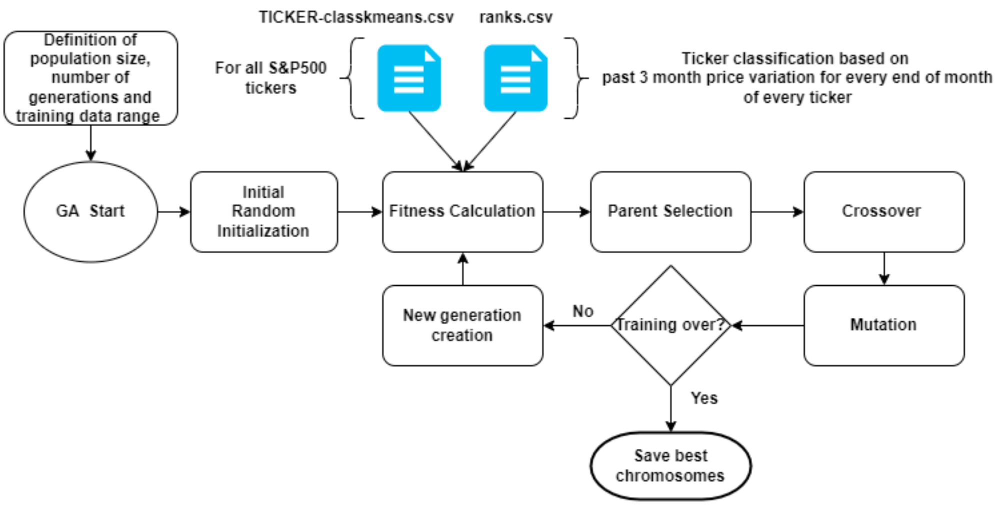

Figure 3.11: Overview of the GA Module

Goal of the GA Module: Broadly speaking, the purpose of this module is to search for fundamental patterns that represent positive price variations in the stock market along various dates. Regarding the fundamental side, for each ticker a classification is done for each end of the month considering every single indicator, generating various files of the type described in figure 3.9. On the other side, for each end of the month, the past 3 month variation of the stock price of each ticker is calculated and used as the criterion for stock classification of that month. An example is provided in figure 3.12.

For date 31-07-2015

Stock ranking example			
Ticker	Past 3 month price variation (%)	Sorted Variation	Final Ranking
MMM	10	40	MRNA
NVDA	-12	10	MMM
MRNA	40	5	SPG
SPG	5	4	TSLA
ZTS	-2	0	GM
TSLA	4	-2	ZTS
GM	0	-12	NVDA

Figure 3.12: Ranking tickers example

In figure 3.13 the ranking method of ticker classification with GA assistance is shown. In this situation, which is an **example**, two fundamental indicators are being considered, FI_1 and FI_2, for two tickers, **AOS** and **BRK.B**, for date **31-07-2015**. The weights WFI_1 and WF_2 represent the weights that a given chromosome is giving to each indicator. This figure illustrates the resulting classifications when combining the weights and the classifications from the indicators which results in a ranking where **AOS** is placed above **BRK.B**. Naturally this is just an **example**, the algorithm considers a significantly larger number of tickers and fundamental indicators and performs this ranking for every single date in the data range. The ranking is also done for the various chromosomes in the population.

Date: 30-09-2020

Ticker	FI_1	FI_2
AOS	12	15
BRK.B	17	13

Weights

WFI_1	WFI_2
0.3	6.4

Overall classification

AOS = FI_1 x WFI_1 + FI_2 x WFI_2 = 12 x 0.3 + 15 x 6.4 = 99.6

BRK.B = FI_1 x WFI_1 + FI_2 x WFI_2 = 17 x 0.3 + 13 x 6.4 = 88.3

Figure 3.13: Example of classification with obtained weights

By comparing the ranking exemplified in figure 3.13 and the price variations ranking exemplified in figure 3.12, this for all dates in the training data range, the algorithm will attempt to find the optimal weights that each fundamental indicator should have such that both rankings are the closest to each other for all dates. The images provided are **merely exemplary** and used to show how both rankings are obtained.

Initial Requirements: Firstly a data range must be set which will correspond to the training set. For example, if the initial date is **31-01-2015** and the final date **31-01-2016** then the periods considered in the training will be all end of the month dates within the range [**31-01-2015**, **31-01-2016**], corresponding to 13 dates in total. As described in 3.11, the type of information required for these dates are the ranks and FIs classifications for each ticker, and their purpose will be described further down when fitness calculation is tackled. The size of the population and number of generations must also be given along with the number of indicators being considered.

Chromosome structure: The chromosome structure will depend entirely on the number of FIs being considered, with each gene representing a ratio. For example, lets assume that N indicators are being considered, then the structure would be:

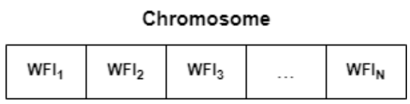

Figure 3.14: Structure of a chromosome

Each gene will represent the weight given to an indicator and, when initializing, a random value between -5 and 5 will be chosen for every gene of all initial population chromosomes.

Fitness Function: As described before, the goal of this module is to find what weights to give to each classified indicator that better mimic the past three month variations at each end of the month of the training set. This way, a fundamental pattern that reflects good stock price variations among various months is the desired objective. The way this is achieved is done by comparing the ranking achieved with the weights and the ranking achieved due to the 3 month price variations, along all training dates. The equation used to calculate the fitness of a chromosome is:

$$Fitness = \sum_{t=1}^{T}\sum_{n=1}^{N}\frac{1}{|OI_{tn} - I_{tn}| + 1}, \qquad \frac{OI_{tn}}{N} < 0.05 \tag{3.1}$$

Where T is the total amount of dates being considered, N the number of tickers available, I_{tn} the index of ticker n in the ranking obtained via GA weights for date t and OI_{tn} the index of ticker n for date t in the ranking generated resorting to the past 3 month variations. It is visible from the equation that the distance between both indexes is the value used when calculating the fitness, closer indexes will affect the fitness positively, whilst bigger differences affect it negatively, considering that the goal is to maximize the fitness. The condition $\frac{OI_{tn}}{N} < 0.05$ indicates that only the top 5% of the original tickers are considered, thus aiming at a more specialized evaluation that is only concerned with the top performers.

Parent Selection/Mating: Half of the current population will be considered for mating with the criteria for parent selection being the chromosome fitness, thus the top fifty percent chromosomes will be chosen for mating.

Crossover: Initially the parents are randomly assigned into pairs that will be responsible for offspring generation. A random probability is given to each pair and, if bigger than the crossover probability p_c (75%), mating will not be considered for said pair. For each set a crossover point c_p is randomly selected and two offspring are created, one that is composed of the first parent until the crossover point and of the second parent until the end. For the other offspring the parents' roles are reversed.

Mutation: The offspring originated by the crossover will suffer mutations in accordance with a probability of mutation p_m (20%). For each gene a percentage is randomly selected and, if bellow p_m, the gene will be altered to a random number within the initialization interval.

After the operators have been applied a new population is created, half composed of the previous generation's best performing chromosomes and the other half composed of their offspring.

Training: When training the GA the population size, number of generations and data ranges must be defined. Firstly the population size selected was 100 with 150 generations. Regarding the training, two years of training for a year of testing was the decision, this due to the fact that since companies' financial statements are only updated each quarter a long term training should be beneficial for pattern finding. Figure 3.15 demonstrates the overall training method, which is the application of a rolling window, thus for each year the GA weights obtained depend on the two years prior. The overall data range considered for training is from **2015** up to **2021**, with testing starting in **2017**

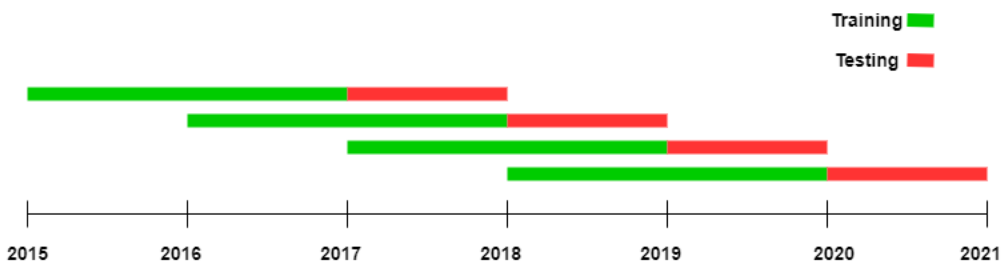

Figure 3.15: Rolling window training

The pseudo-code for the GA model is:

Algorithm 3.4: GA Module Program - Pseudo-code

```
Load the past three month variations for each end of the month of each ticker considered
Obtain the classifications per ticker calculated in the Stock Classifier Module for a specific
  strategy
Define the training data range
function init_ga(ranks_by_date, classification_per_ticker, init_date, end_date)
    Define the population size, number of generations and number of parents mating
    Initialize the population
    for each generation:
        Calculate the fitness for the entire population
        Select the mating pool based on fitness
        Generate the next generation using crossover
        Add some variations to the offspring using mutation
        Create the new population based on the parents and offspring
    Return the final population obtained
Save the best solution
```

3.6 Portfolio Module

The Portfolio Module, described in figure 3.16, is responsible for the selection and testing of various portfolios based on the weights obtained in the GA Module for the range of **01-02-2017** - **31-10-2021**.

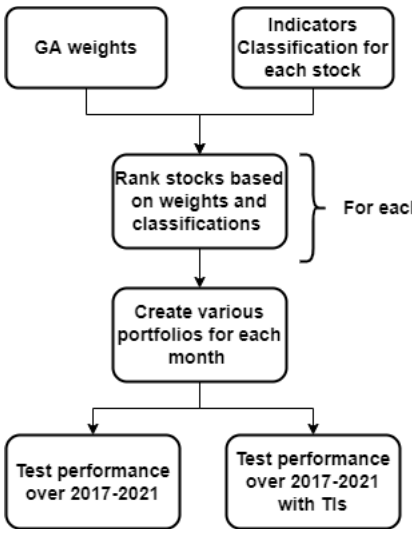

Figure 3.16: Overall view of the Portfolio Module

Initially, in this module, a python program receives the weights from the **GA Module** and the classifications obtained in the **Stock Classifier Module**. Utilizing these two it is possible to generate a new more nuanced classification that takes into account the importance of each indicator. Taking advantage of this new ranking sets of stocks are selected to generate portfolios. The four types of sets considered are: The best two stocks, the best five stocks, the best seven stocks and finally the best stock for each sector. It is important to reference that, following the way training proceeds these weights obtained will be different every year. For each strategy, **Growth Strategy**, **Balance Strategy** and **Value Strategy**, the algorithm was run four times, totalling twelve sets of chromosomes that will be used to form portfolios, four for each strategy. Each set will have five different weight combinations, one for each year.

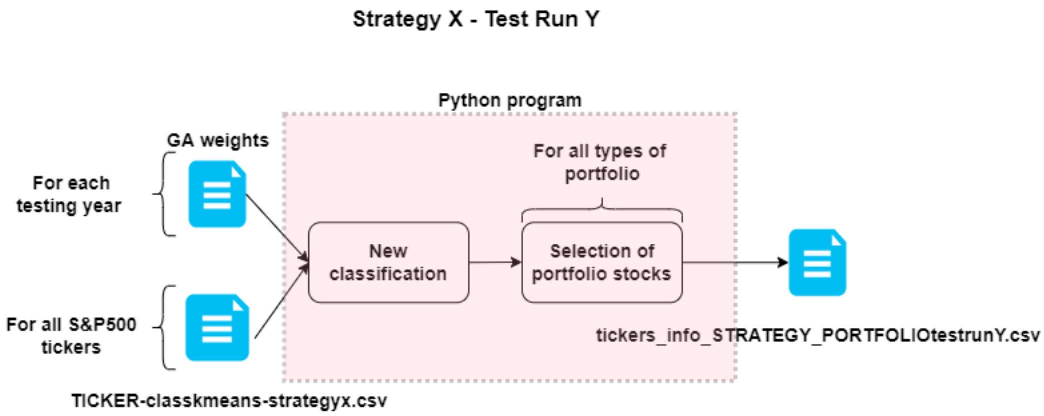

Figure 3.17: Portfolio Generation

In figure 3.17, the method for generating portfolios is briefly explained. Since the schematic is **merely exemplary**, Strategy X may be any of the strategies defined in the **Input Strategy** and Test Run Y may be any of the test runs explained in the paragraph above. Firstly, the program will receive the GA weights for each year (2017-2021) for Strategy X and the classifications of each S&P 500 stock corresponding to Strategy X from the Stock Classifier. Note that the GA weights for each testing year were obtained via the training done in the **two years** prior. With these files the program will generate a new classification for each end of month which will be used to pick stocks for the next month of the portfolio, by picking the best newly classified stocks. After this is done a file, *tickers_info_STRATEGY_PORTFOLIOtestrunY*.csv, is created corresponding to the stocks used to create the portfolio for test run *Y*, with the type of portfolio *PORTFOLIO* (2, 5, 7 or Sectors stocks) when following the strategy *STRATEGY* (Growth, Balanced or Value), this for all months considered in testing. The structure of this file is given in 3.18

Since the classifier updates the classifications of stocks at the end of each month the new ranking must also be updated with the same frequency. So this file will store the best tickers at the end of the month defined in the column **end_calc** and the period of trading will correspond to the interval between **trading_start** and **trading_end**. The tickers, depending on the portfolio, belong to columns **ticker1** through **tickerN** and, at the end of the 31st of January of each year the GA weights considered are updated as per the method of training of the GA. The pseudo-code of the Portfolio Generation program is shown in 3.5.

end_calc	trading_start	trading_end	ticker1	...	...	tickerN
2017-1-31	2017-2-01	2017-2-28	MKTX	...	...	FAST
...	...	...	...	...	...	...
2018-1-31	2018-2-01	2018-2-28	NEE	...	...	ABMD
...	...	...	...	...	...	...
2019-1-31	2019-2-01	2019-2-28	MNST	...	...	CTRA
...	...	...	...	...	...	...
2020-1-31	2020-2-01	2020-2-29	ETSY	...	...	MNST
...	...	...	...	...	...	...
2021-1-31	2021-2-01	2021-2-28	PPL	...	...	TWTR
...	...	...	...	...	...	...
2021-9-30	2021-10-01	2021-10-31	AES	...	...	WYNN

Figure 3.18: File storing the tickers that will compose the portfolio for each month (Tickers used are exemplary)

Algorithm 3.5: Portfolio Generation Program - Pseudo-code

Open the classifications obtained in the Stock Classifier Module for the training/testing period
Access the GA weights obtained for each training year
Access the tickers being used and their respective sectors
function new_classification(classifications_per_ticker, tickers_and_sectors, ga_weights, dates)
    Define testing data range (every end of the month involved)
    Verify current year
    while inside the data range:
        function calc_class_weights(classifications_per_ticker, ga_weights[current_year], current_date)
            Considering the current year and date reclassify the stocks using the GA weights for each indicator
            return the newly obtained classifications
        Save the tickers used for the creation of the portfolio for that particular trading month
        Save the end_calc, trading_start and trading_end for that particular trading month
    return the chosen stocks with respective dates for the testing period
Save information in a csv file like the one in figure 3.18

This file will then be used by another python program in order to test the performance of the entire model. Two different methods will be applied, a *laissez-faire* method that merely generates equally distributed portfolios and a more controlling method that, although still applying equal distribution, takes advantage of a simple technical indicator to define entry and exit points for the portfolio stocks each month. The indicator used will be the EMA considering a 24-day window, which was the window that delivered the best results. The application of this indicator is simple, if the EMA of the previous day is bellow today's price a BUY signal is sent for that particular stock, else a SELL signal will be registered, as illustrated by 3.19.

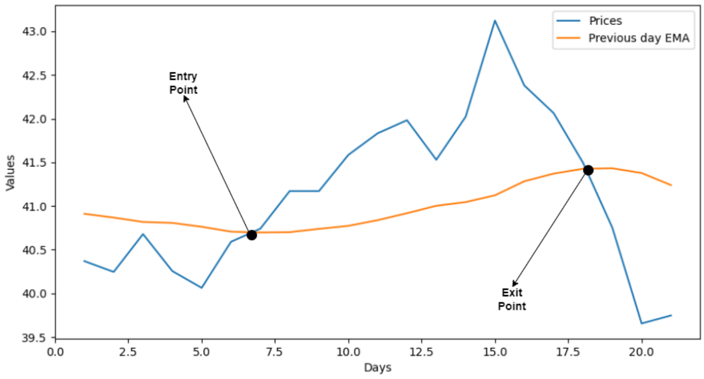

Figure 3.19: Application of the EMA trading system for one month (Purely exemplary)

The pseudo-code for the portfolio testing program is:

Algorithm 3.6: Portfolio Testing Program - Pseudo-code

1: Open each portfolio file obtained in 3.5 for a particular portfolio type and strategy
2: **for** each file (test run):
3: **function** load_portfolio(file)
4 **return** the file in the form of a dictionary
5: Create a list that includes each one of these dictionaries
6: **for** each dictionary on the list:
7: Define the starting money for the portfolio (with and without trading system)
8: **for** each trading month in the dictionary:
9: Register current end_calc, trading_start and trading_end
10: Check and register if current stocks were traded in the previous trading period (presence signals)
11: **function** obtain_returns(end_calc, trading_start, trading_end, stocks, starting_money, presence_signals, strategy, prev_position)
12: Obtain the stock variations for the companies in the defined data range
13: **if** the trading system is applied obtain the trading flags (IN/OUT) for each stock in the portfolio for the trading period
14: Calculate the portfolio money obtained after the trading is done for that specific trading period, considering commissions
15: **return** the updated money, the trading positions and the number of trades
16: Update the starting money, previous positions and number of trades
17: Store the monthly performance of that specific portfolio with and without the trading system
18: Store the overall portfolio performances for all portfolios over all test runs
19: Plot the results across the entire testing period and compare them to the benchmark

4

Results

Contents

In this chapter various different experiments are done in order to test the model proposed in this thesis. Multiple evaluation metrics are used to assess results, alongside a benchmark.

4.1 Introduction

The GA described in the section above was tested for years 2017 to 2021. For each year the algorithm was run various times in order to get a better understanding of its general performance. It is important to note that **the labeling of test runs is purely for organizational purposes and to get a better understanding of a long term performance, i.e, there is no correlation between Test Run 1 2017 and Test Run 1 2018 and so on**, thus, any random shuffle of Test Runs is valid for a long term (2017-2021) observation, the labeling of Test Runs is merely for simplicity's sake.

As stated before, three different strategies were tested, the **Growth Strategy**, where the algorithm will analyse growth focused FIs; the **Balanced Strategy**, which will take into consideration a wider range of indicators and the **Value Strategy**, that will attempt to prioritize the most stable companies. For the **Growth Strategy** five indicators were considered, D/E, ROE, NPM, RG and RR; for the **Balanced Strategy** nine indicators were considered, D/E, CR, EPS, ROE, ROA, NPM, GPM, RG and OM and finally for the **Value Strategy** five indicators were considered, CR, EPS, ROA, RG, PR.

In regards to the portfolio compositions, various different compositions were considered, a 2 Stock Portfolio (2SP), picking the top two companies for each month, a 5 Stock Portfolio (5SP), a 7 Stock Portfolio (7SP), both following the same logic as the 2SP, and a Sector Stock Portfolio (SSP), where a company from each sector is considered. The reason these are the proposed compositions is to analyse how the GA performs when taking into account variance, since naturally smaller portfolios are considered less stable, allowing for better performances but also greater failures.

These portfolios were all updated monthly according to the combination of classifier + GA. For all test runs two situations were considered, one where, for each month, the portfolio consists of the equal distribution of the stocks selected and another where, although the investment is equally distributed among stocks, a EMA based trading system is implemented to decide entry/exit points from the market.

The initial investment considered was of **10000**, the period considered for the EMA was **24 days**, the commission considered was **2%** of the money invested upon stock purchase and the GA features were as described in table 4.1.

In order to ascertain the performance of the model four experiments were conducted, analysing and comparing the results from all three strategies over the four different portfolio types described above.

Table 4.1: GA features

Number of Generations	150
Population Size	100
Number of Parents Mating	50
Crossover Probability	0.75
Mutation Probability	0.20

4.2 Evaluation Metrics

4.2.1 Return on Investment (Return On Investment (ROI))

The ROI is a measure that allows an investor to understand the profitability of a particular investment, specially when comparing to other investments. The basic equation of the ROI is:

$$ROI = \frac{ValueofInvestment - CostofInvestment}{CostofInvestment} \tag{4.1}$$

It is important to note that this metric is best when used to compare different strategies so, for all experiments, all Test Runs will be compared to the benchmark (S&P 500) and to each other. Since the returns presented in all experiments are yearly and the portfolios are updated monthly the ROI observed differs slightly from equation 4.1, with the money generated from previous months being reinvested.

$$ROI = \frac{ValueofInvestment - ReturnsofFormerInvestment}{ReturnsofFormerInvestment} \tag{4.2}$$

Finally, all yearly returns are compounded to analyse the long term performance.

4.2.2 Sharpe Ratio

The Sharpe Ratio (SR) is used in finance to evaluate the risk/return of an investment. It provides a perception to the investor of how much more risk is being taken for higher returns.

$$SharpeRatio(P) = \frac{R_P - R_F}{\sigma_P} \tag{4.3}$$

Where R_P is the average rate of return of the portfolio P, R_F is the risk-free rate and σ_P is the portfolio P's standard deviation. In this thesis a 0 risk-free rate is considered due to currently low interest rates.

4.2.3 Positive Month Percentage (PMP)

Like the name indicates this metric will measure the percentage of months where a particular Test Run had a positive increase in value. This will be calculated for the entire testing period.

$$\text{PMP} = \frac{Number of months above zero}{Total number of trading months} \tag{4.4}$$

4.2.4 Benchmark Beating Percentage (BBP)

The BBP will register the percentage of months were the portfolio was able to outperform the market.

$$\text{BBP} = \frac{Number of months above the benchmark}{Total number of trading months} \tag{4.5}$$

4.2.5 Number of trades

The number of times a stock in the portfolio is sold or bought will be registered for each Test Run, this for the entire duration of testing. When considering the equal distribution portfolios without the trading system, since the portfolio is updated monthly, at the end of every month/beginning of the following month the number of entries/exits is registered as two trades, the selling of the previous stock and the purchase of the new stock. Taking the trading system into consideration, every time a stock is bought/sold during a particular month due to signals given by the EMA system a trade will be registered.

4.3 Experiment I - 2 Stock Portfolio

Table 4.2 shows the yearly as well as total ROIs obtained for all Test Runs and all strategies when considering the 2SP .

Table 4.2: ROI for all strategies for the 2SP including total compound returns

Strategy	Test Run	2017	2018	2019	2020	2021	Total
Growth Strategy	1	48.1%	-13.8%	32.7%	-4.4%	111.0%	241.5%
	1 EMA	46.9%	0.0%	24.9%	8.4%	82.8%	263.6%
	2	38.1%	-36.1%	14.6%	23.6%	-16.5%	4.5%
	2 EMA	31.8%	-0.7%	10.2%	43.3%	-8.3%	89.7%
	3	50.7%	-35.5%	25.3%	18.1%	28.4%	84.9%
	3 EMA	49.8%	0.3%	22.0%	29.0%	24.4%	194.2%
	4	52.0%	-20.6%	-8.6%	-7.1%	-7.6%	-7.3%
	4 EMA	53.3%	6.5%	9.9%	-2.9%	-2.4%	69.9%
Balanced Strategy	1	60.4%	-6.8%	22.1%	102.7%	59.0%	488.5%
	1 EMA	44.0%	26.8%	33.2%	119.4%	22.8%	555.2%
	2	10.8%	-8.9%	-2.6%	75.6%	18.7%	105.1%
	2 EMA	5.0%	19.5%	7.3%	82.6%	15.6%	184.4%
	3	52.4%	3.8%	7.5%	57.7%	8.3%	190.5%
	3 EMA	20.2%	14.0%	5.8%	81.4%	-0.4%	161.9%
	4	45.8%	14.5%	18.7%	22.0%	17.5%	184.1%
	4 EMA	13.1%	26.8%	38.8%	94.8%	17.6%	356.1%
Value Strategy	1	46.7%	34.5%	52.6%	65.3%	13.8%	466.6%
	1 EMA	13.7%	47.7%	40.0%	62.7%	36.1%	420.5%
	2	32.1%	10.3%	16.0%	100.2%	17.8%	298.5%
	2 EMA	17.6%	20.6%	10.3%	89.5%	16.2%	244.6%
	3	52.0%	6.6%	52.6%	84.0%	49.4%	579.7%
	3 EMA	28.0%	11.9%	40.0%	74.2%	48.3%	418.0%
	4	52.0%	-18.6%	21.1%	111.1%	9.1%	245.4%
	4 EMA	26.7%	-0.8%	12.6%	88.9%	12.4%	200.5%
S&P 500	-	15.8%	-7.3%	23.2%	16.8%	18.5%	83.0%

For the **Growth Strategy** the average compound ROI of all Test Runs, with and without the EMA strategy was 117.6%, which was 34.6 percentage points above the index. It is important to note the instability of the performances, with the gap between best and worst performance being 270.9 percentage points and three Test Runs being bellow the benchmark. The introduction of the EMA trading system improved the returns of all Test Runs, playing a especially important role in **2018** and **2020**. Figure 4.1 plots the various Test Runs for the **Growth Strategy**.

Regarding the **Balanced Strategy** an improvement in results was verified, with the average compound returns being 278.2%, more than 3 times better than the S&P 500. Although the gap between best and worst Test Run was of 450.1 percentage points, higher than that of the **Growth Strategy**, all performances were above the index. Barring **Test Run 3**, all Test Runs benefited from the introduction of the EMA trading system. Figure 4.2 plots the various Test Runs for the **Balanced Strategy**.

Finally, for the **Value Strategy** the average compound ROI was of 359.2%, the best of all strategies. Comparing the best and worst Test Runs the difference was of 379.2 percentage points, although the worst Test Run was already 2.4 times better than the market. For this particular strategy the effect of

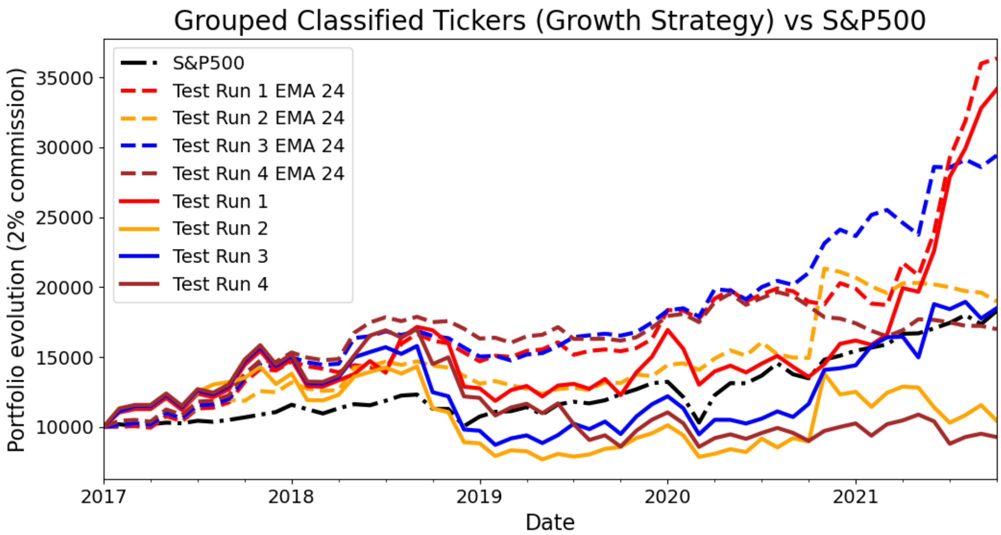

Figure 4.1: 2 Stock Portfolio evolution (Growth) vs the S&P500 (y-axis - €)

the trading system was diminished, with its introduction being unable to improve any Test Runs, granted that it didn't severely hinder any particular Test Run either. Figure 4.3 plots the various Test Runs for the **Value Strategy**.

Overall the best strategy for this portfolio was the **Value Strategy**, followed up by the **Balanced Strategy** and the **Growth Strategy**. Given that the portfolio composition considered was the 2SP the differences between best and worst Test Run can be justified by high volatility, making the results inconsistent, especially when paired with the **Growth Strategy** that already trends towards unstable companies. The EMA trading system proved to be a useful tool, particularly when avoiding losses for more unstable strategies.

Figure 4.1 shows the long term performances of all Test Runs for the **Growth Strategy**. It is clear that, up until around September of 2018 where the market downtrend happened, all Test Runs were performing quite well. The dip that occurred between September and December of that year highly influenced Test Run returns with some Test Runs never recovering in subsequent years. For the **Balanced Strategy**, as figure 4.2 indicates, although the market downtrends inflicted some losses for all Test Runs the recovery was much better with 2020 and 2021 being particularly good years. Figure 4.3 displays the strength of the **Value Strategy** which was able to consistently yield results throughout the years for most Test Runs, with the returns exploding in 2020, indicating that this strategy was extremely strong during big market downtrends and subsequent rises.

As table 4.3 shows the strategy that had the best SR values for all its Test Runs was the **Value Strategy**, indicating that the best strategy proved to also be the less risky. The **Balanced Strategy** was also considered the riskiest strategy, closely followed by the **Growth Strategy**.

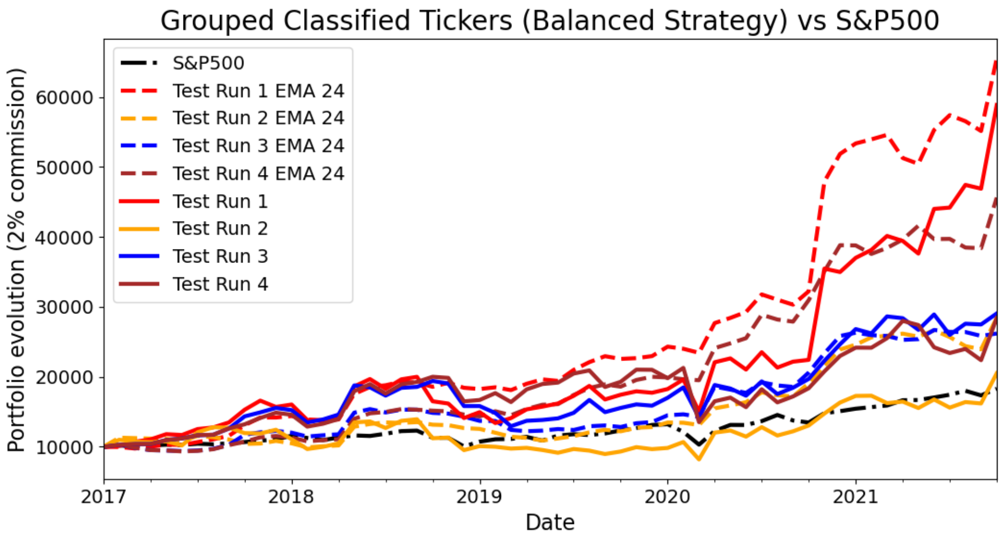

Figure 4.2: 2 Stock Portfolio (Balanced) evolution vs the S&P500 (y-axis - €)

4.4 Experiment II - 5 Stock Portfolio

Table 4.4 shows the yearly as well as total ROIs obtained for all Test Runs and all strategies when considering the 5SP .

When considering the 5SP, the **Growth Strategy** had an average compound ROI of 107.2%, slightly above the market. Comparing to its 2SP counterpart the average returns were reduced by 10.4 percentage points, possibly because the introduction of more stocks minimized volatility which will make extremely positive/negative outliers less common. Regarding the gap between best and worst performance, for the 5SP it was of 133.7 percentage points, strengthening the idea that the increase in stocks tends towards more stable results. The EMA trading system was again an indispensable tool, greatly improving all Test Runs. Figure 4.4 plots the various Test Runs for the **Growth Strategy**.

For the **Balanced Strategy** the average compound ROI was of 210.9%, also seeing a reduction in returns when compared to the 2SP. In this case all Test Runs barring **Test Run 4 EMA** performed worst that the 2SP Test Runs, showing that the introduction of stocks lacked an upside for this particular strategy. The difference between worst and best Test Run was 187.1 percentage points so, as expected, the results became more stable at the cost of performance. The impact of the trading system was still mostly positive. Figure 4.5 plots the various Test Runs for the **Balanced Strategy**.

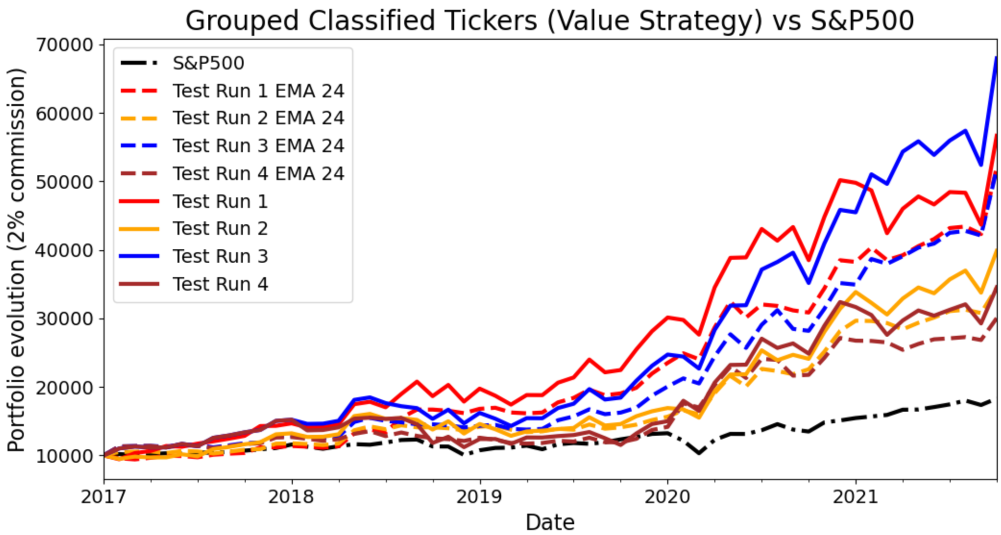

Figure 4.3: 2 Stock Portfolio (Value) evolution vs the S&P500 (y-axis - €)

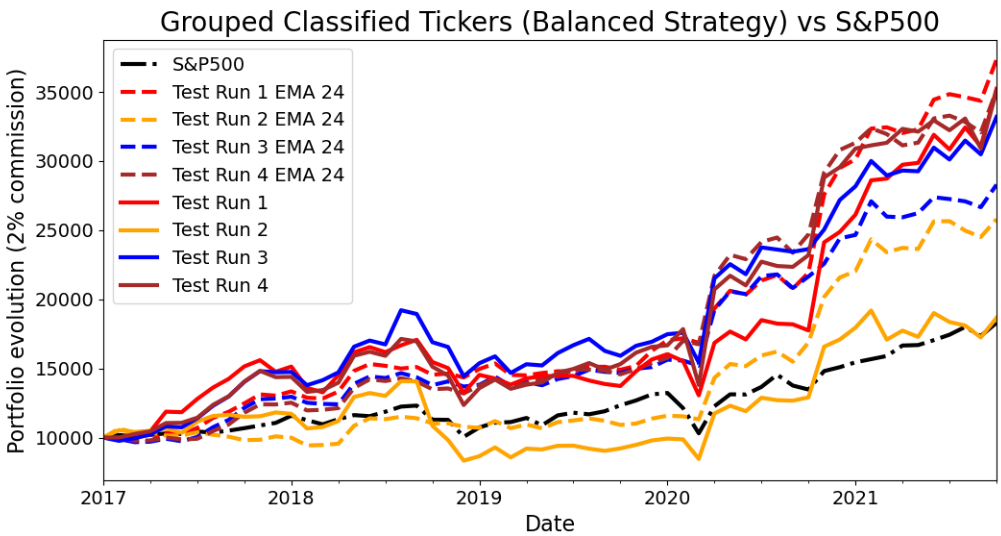

Figure 4.5: 5 Stock Portfolio evolution (Balanced) vs the S&P500 (y-axis - €)

Regarding the **Value Strategy**, the average compound ROI was 243.8%, still being the best among strategies. A similar trend could be verified with the addition of stocks hindering returns but also improving the stability between Test Runs, with the difference between best and worst Test Run being 161.8 percentage points. Interestingly enough, unlike the 2SP, the 5SP actually benefited from the introduction of the EMA trading system for this particular strategy. Figure 4.6 plots the various Test Runs for the **Value Strategy**.

Table 4.3: BBP, PMP, number of trades and SR for strategies for the 2SP

Strategy	Test Run	BBP	PMP	Number of trades	SR
Growth Strategy	1	52.632%	61.404%	86	0.69
	1 EMA	49.123%	56.140%	289	0.69
	2	45.614%	56.140%	92	0.64
	2 EMA	40.351%	45.614%	276	0.64
	3	45.614%	59.694%	90	0.64
	3 EMA	52.632%	63.158%	269	0.64
	4	49.123%	57.895%	84	0.62
	4 EMA	45.614%	52.632%	273	0.62
Balanced Strategy	1	59.649%	68.421%	88	0.64
	1 EMA	56.140%	63.158%	263	0.64
	2	50.877%	56.140%	88	0.62
	2 EMA	50.877%	52.632%	316	0.62
	3	61.404%	64.912%	106	0.64
	3 EMA	42.105%	50.877%	319	0.64
	4	59.649%	70.175%	98	0.63
	4 EMA	50.877%	59.649%	277	0.63
Value Strategy	1	63.158%	64.912%	82	0.72
	1 EMA	61.404%	63.158%	258	0.72
	2	57.895%	66.667%	70	0.77
	2 EMA	56.140%	61.404%	264	0.77
	3	61.404%	66.667%	84	0.78
	3 EMA	57.895%	63.158%	263	0.78
	4	56.140%	66.667%	80	0.79
	4 EMA	52.632%	56.140%	282	0.79

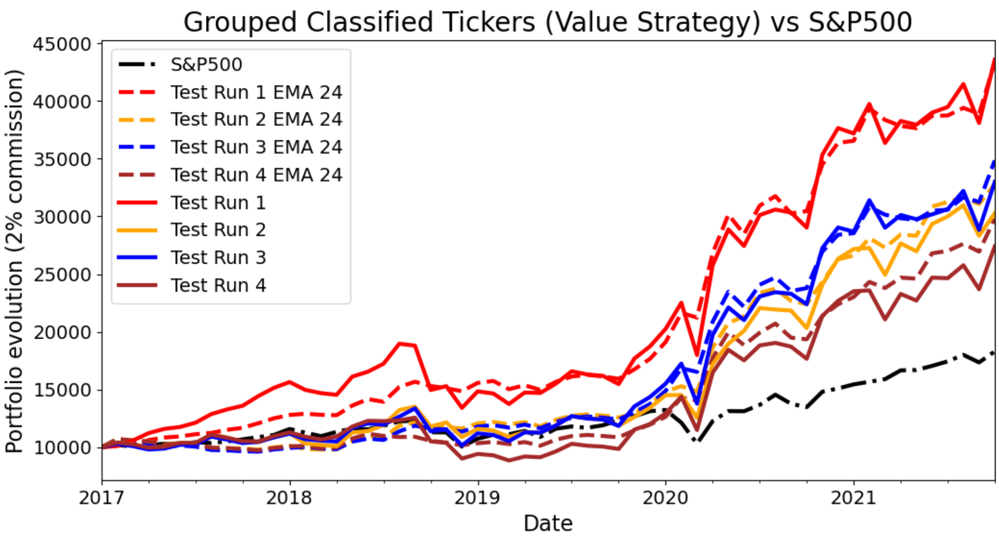

Figure 4.6: 5 Stock Portfolio evolution (Value) vs the S&P500 (y-axis - €)

Overall the sequence of strategy performances remained the same, with the **Value Strategy** still

Table 4.4: ROI for all strategies for the 5SP including total compound returns

Strategy	Test Run	2017	2018	2019	2020	2021	Total
Growth Strategy	1	28.7%	-14.6%	18.3%	-12.8%	54.9%	75.7%
	1 EMA	25.5%	-0.7%	17.3%	14.9%	40.7%	136.3%
	2	25.9%	-8.1%	10.3%	2.2%	-0.2%	30.0%
	2 EMA	27.4%	3.2%	7.6%	20.2%	8.1%	83.9%
	3	45.0 %	-6.0 %	21.3%	-12.7%	71.0%	146.8%
	3 EMA	33.2%	4.2%	17.8%	11.7%	44.3%	163.7%
	4	36.9%	-0.9%	22.2%	-9.3%	27.4%	91.7%
	4 EMA	29.5%	19.4%	17.1%	1.3%	25.0%	129.3%
Balanced Strategy	1	51.3%	-4.0%	10.3%	63.0%	34.8%	252.2%
	1 EMA	33.4%	12.1%	13.8%	76.9%	24.0%	273.7%
	2	17.2%	-26.1%	14.7%	80.8%	3.9%	86.6%
	2 EMA	-0.1%	7.2%	8.2%	90.2%	17.0%	158.0%
	3	47.5%	4.3%	13.6%	61.2%	17.8%	231.9%
	3 EMA	29.5%	6.7%	13.2%	57.7%	14.7%	182.9%
	4	43.9%	-5.7%	22.8%	85.4%	13.4%	250.5%
	4 EMA	25.2%	9.9%	14.6%	98.7%	12.1%	251.2%
Value Strategy	1	56.5%	-5.2%	36.5%	83.6%	17.2%	335.7%
	1 EMA	28.0%	21.7%	22.8%	91.2%	18.6%	333.7%
	2	12.1%	3.4%	25.2%	87.3%	11.4%	202.7%
	2 EMA	1.4%	19.0%	23.0%	79.4%	23.2%	227.8%
	3	11.8%	0.2%	38.5%	84.8%	15.0%	229.8%
	3 EMA	-0.6%	18.5%	26.3%	91.8%	22.0%	248.40%
	4	12.9%	-16.6%	37.4%	81.7%	16.5%	173.9%
	4 EMA	0.7%	2.7%	22.2%	82.0%	29.5%	198.1%
S&P 500	-	15.8%	-7.3%	23.2%	16.8%	18.5%	83.0%

being the best, followed by the **Balanced Strategy** and then the **Growth Strategy**. The introduction of more stocks saw improvements in stability but reduction in returns, specially when considering the best performing Test Runs. Given that the **Value Strategy** and **Balanced Strategy** had already shown promising results on average for the 2SP, the 5SP appears to have a mostly negative impact. A different story can be told for the **Growth Strategy** since, although the best Test Runs had diminished returns, given the instability of the 2SP, a more risk averse investor will benefit from the improvement in stability, especially the increase of returns for the worst Test Runs, which at the very least avoids major losses.

The **Growth Strategy** presented, following 4.4, overall good returns up until the end of 2018. After that some Test Runs, especially when applying the EMA trading system, were able to recover, whilst others struggled. Overall it appears that this strategy had a hard time during and after market dips. Figure 4.5 shows that most Test Runs, for the **Balanced Strategy**, were able to beat the market throughout the years, with 2017 and 2020 being the most profitable years. The EMA system clearly avoids heavy losses but also hinders great returns. The **Value Strategy** proved to be the most consistent strategy again, as figure 4.6 indicates, although most Test Runs were only able to find great returns after 2020,

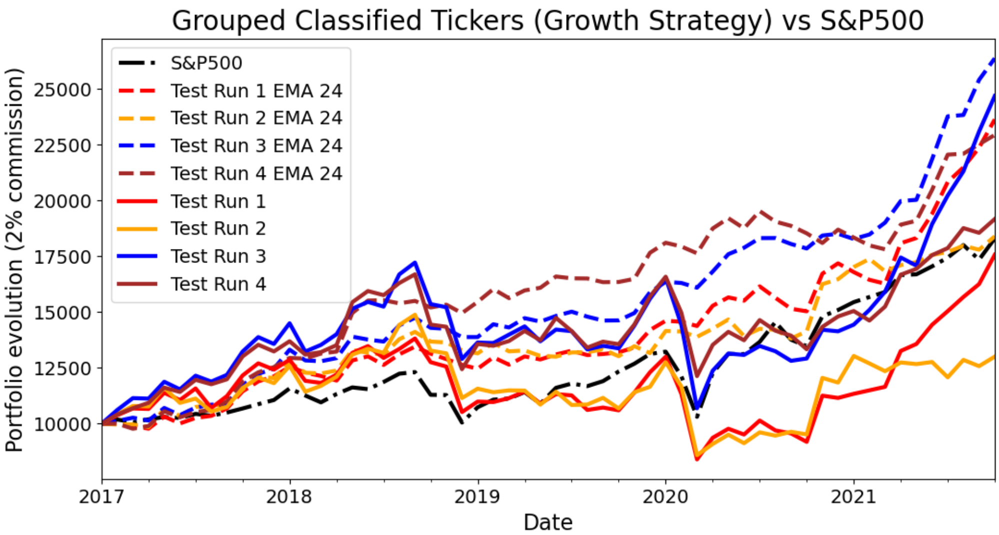

Figure 4.4: 5 Stock Portfolio evolution (Growth) vs the S&P500 (y-axis - €)

consolidating the idea that this strategy mostly benefits from accentuated market declines.

Table 4.5 remained consistent with table 4.3, with the best SR strategy still being the **Value Strategy**, followed by the **Growth Strategy** and **Balanced Strategy**. It is interesting to note that the increase in stocks actually reduced the SR, opposite to the intuitive expected increase.

4.5 Experiment III - 7 Stock Portfolio

Table 4.6 shows the yearly as well as total ROIs obtained for all Test Runs and all strategies when considering the 7SP .

The **Growth Strategy** had, for the 7SP, 110.1% average compound returns, better than the 5SP. The worst Test Run for this type of portfolio was better that the worst Test Runs for the 2SP and 5SP and the best Test Run was worse than its 2SP and 7SP, indicating that the addition of stocks led to a more stable position. Regarding the difference between best and worst Test Run it was 119.2 percentage points. In this case the Test Runs mostly outperformed the S&P 500, especially when the trading system was applied. Until now this portfolio seems to be the best suited for a risk averse investor that is considering the **Growth Strategy**. Figure 4.7 plots the various Test Runs for the **Growth Strategy**.

A slight improvement of the worst Test Run and a slight general dip in performance for the best Test Runs was found for the **Balanced Strategy**, averaging 196.9% compound returns. Although a general reduction in average returns was found the gap between worst and best Test Run was the smallest, being 125.1 percentage points and the worst Test Run had the best performance of all the **Balanced Strategy**'s Test Runs. The results obtained when using the EMA trading system didn't deviate much

Table 4.5: BBP, PMP, number of trades and SR for strategies for the 5SP

Strategy	Test Run	BBP	PMP	Number of trades	SR
Growth Strategy	1	45.614%	61.403%	235	0.65
	1 EMA	57.895%	59.649%	754	0.65
	2	43.860%	56.140%	229	0.62
	2 EMA	47.368%	56.140%	714	0.62
	3	52.632%	63.158%	223	0.60
	3 EMA	52.632%	68.421%	707	0.60
	4	54.386%	63.158%	203	0.62
	4 EMA	52.632%	64.912%	724	0.62
Balanced Strategy	1	56.140%	59.649%	181	0.63
	1 EMA	52.632%	66.667%	712	0.63
	2	45.614%	57.895%	175	0.62
	2 EMA	43.860%	56.140%	759	0.62
	3	57.895%	63.158%	175	0.60
	3 EMA	49.122%	59.649%	709	0.60
	4	52.632%	71.930%	193	0.60
	4 EMA	50.877%	66.667%	718	0.60
Value Strategy	Test Run 1	61.404%	64.912%	159	0.70
	1 EMA	61.404%	68.421%	665	0.70
	2	54.386%	61.404%	171	0.68
	2 EMA	49.123%	57.895%	699	0.68
	3	45.614%	57.895%	175	0.69
	3 EMA	54.386%	56.140%	701	0.69
	4	45.614%	56.140%	173	0.70
	4 EMA	49.123%	56.140%	724	0.70

from its equal distribution counterparts, with some Test Runs improving returns and others reducing them. Figure 4.8 plots the various Test Runs for the **Balanced Strategy**.

For the **Value Strategy** the average compound returns were 202.3%, the best among all strategies. The addition of stocks had, overall, a negative effect on the strategy, with both worst and best Test Runs seeing a reduction in returns. It appears that, contrary to other strategies where there is more room for debate, the increase in stocks negatively impacts the **Value Strategy**, highlighting the capability of the algorithm to find companies that will perform even when considering extremely volatile portfolios. Figure 4.9 plots the various Test Runs for the **Value Strategy**.

Up until now the **Value Strategy** stands out as the best strategy for all portfolios, not only yielding the best returns but also always outperforming the market. As the number of stocks increases the **Balanced Strategy** gets closer to matching the **Value Strategy**, with the number of stocks ostensibly having opposite effects on these strategies. The **Growth Strategy** also saw improvements with the introduction of more stocks but it is still the clear least viable strategy.

In this case the **Growth Strategy** was able to withstand the 2018 market decline, with all Test Runs staying above the benchmark. As figure 4.7 shows, this wasn't true for 2020, which registered major

Table 4.6: ROI for all strategies for the 7SP including total compound returns

Strategy	Test Run	2017	2018	2019	2020	2021	Total
Growth Strategy	1	43.8%	-20.3%	31.4%	-14.3%	40.8%	81.8%
	1 EMA	31.5%	1.5%	24.5%	16.2%	30.1%	151.2%
	2	22.9 %	-4.1 %	16.9%	1.0 %	-0.8%	38.1%
	2 EMA	26.3%	6.9%	10.7%	13.6%	3.0%	75.0%
	3	40.8%	-3.6%	28.0%	-5.1%	46.6%	141.7%
	3 EMA	31.9%	4.9%	25.5%	9.1%	35.8%	157.3%
	4	39.1%	-1.2%	35.9%	-8.3%	24.0%	112.4%
	4 EMA	19.5%	12.9%	26.6%	12.9%	15.8%	123.4%
Balanced Strategy	1	56.4%	-8.3%	13.5%	65.5%	22.3%	229.4%
	1 EMA	37.1%	8.7%	9.4%	67.7%	18.5%	224.1%
	2	12.3%	-26.7%	27.9%	94.2%	10.1 %	124.9%
	2 EMA	-3.6%	4.1%	14.3%	84.6%	18.9%	151.9%
	3	45.7 %	-5.7%	16.0%	64.9%	15.3%	202.9%
	3 EMA	26.2%	11.6%	9.9%	57.0%	12.8%	173.9%
	4	47.2 %	-15.1 %	24.1 %	74.9 %	17.4%	218.4%
	4 EMA	27.3%	8.4%	14.8%	89.2%	16.8%	250.0%
Value Strategy	1	34.0%	0.1%	31.2%	56.5%	19.0%	227.9%
	1 EMA	17.5%	17.4%	21.1%	97.9%	20.1%	296.8%
	2	10.9%	-8.4%	36.4%	73.4%	8.8%	161.6%
	2 EMA	1.1%	13.7%	21.8%	81.5%	17.1%	197.8%
	3	16.5%	-0.1%	31.2%	56.5%	18.1%	182.0%
	3 EMA	4.2%	15.3%	21.0%	97.3%	18.6%	240.4%
	4	18.0%	-15.8%	32.4%	57.7%	13.6%	135.5%
	4 EMA	4.1%	5.5%	22.4%	71.3%	20.2%	176.7%
S&P 500	-	15.8%	-7.3%	23.2%	16.8%	18.5%	83.0%

losses during the market decline, especially when the trading system wasn't involved. Figure 4.8 also shows how most Test Runs were able to stay above the benchmark for the duration of the **Balanced Strategy** testing, with the continuous growth being best before the decline in 2018 and after the downtrend in 2020. The **Value Strategy**, as shown by figure 4.9, somewhat followed the benchmark up until 2020, where all Test Runs started obtaining great returns. Overall it appears that value companies are the best ones to invest in during big market downtrends.

The Sharpe Ratios when considering the 7SP remained similar to or slightly bellow those of the 5SP. As table 4.7 indicates the best strategy was still the **Value Strategy** and the worst strategy the **Balanced Strategy**.

4.6 Experiment IV - Sectors Stock Portfolio

Table 4.8 shows the yearly as well as total ROIs obtained for all Test Runs and all strategies when considering the SSP .

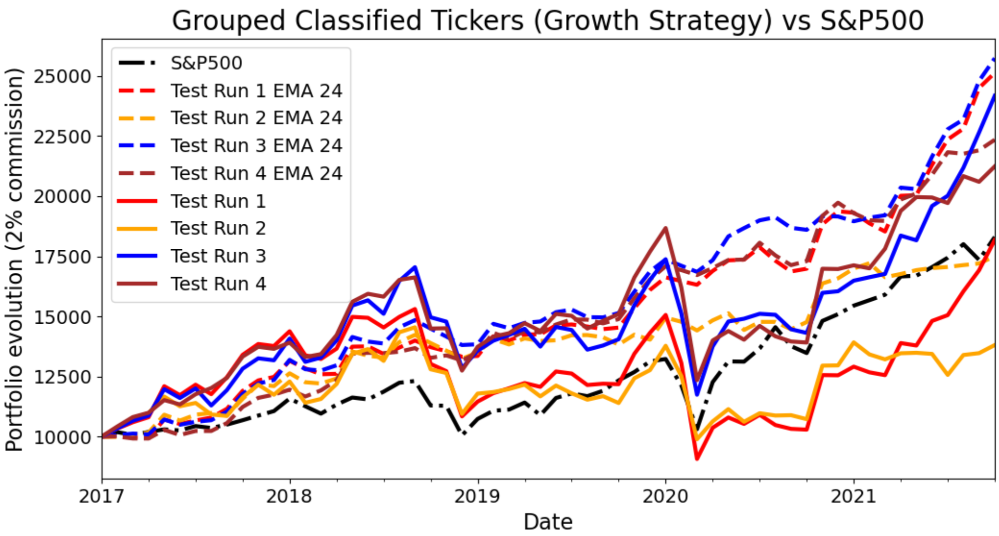

Figure 4.7: 7 Stock Portfolio evolution (Growth) vs the S&P500 (y-axis - €)

When using the SSP the **Growth Strategy** averaged 91.5% compound returns, extremely close to the market returns. The gap between the worst and best Test Run was also 20.4 percentage points, indicating this portfolio was the most stable portfolio and more or less mimics the S&P 500, with most Test Runs being slightly above it. The introduction of the EMA trading system has very little impact on performances. Figure 4.10 plots the various Test Runs for the **Growth Strategy**.

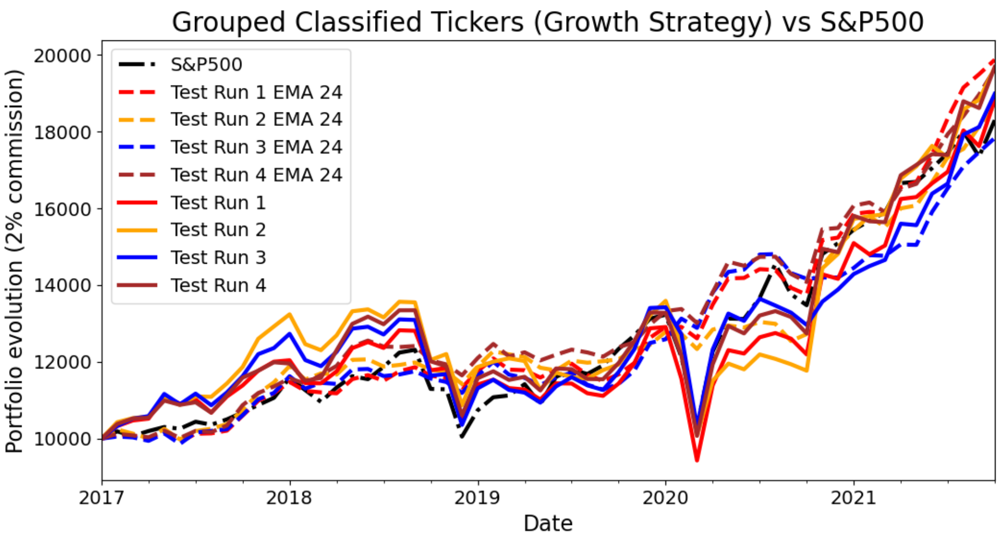

Figure 4.10: Sector Stock Portfolio evolution vs the S&P500 (y-axis - €)

The **Balanced Strategy** got 142.0% average compound ROI indicating that the addition of stocks and sector distribution influenced the average performance negatively. Comparing to the 7SP both the

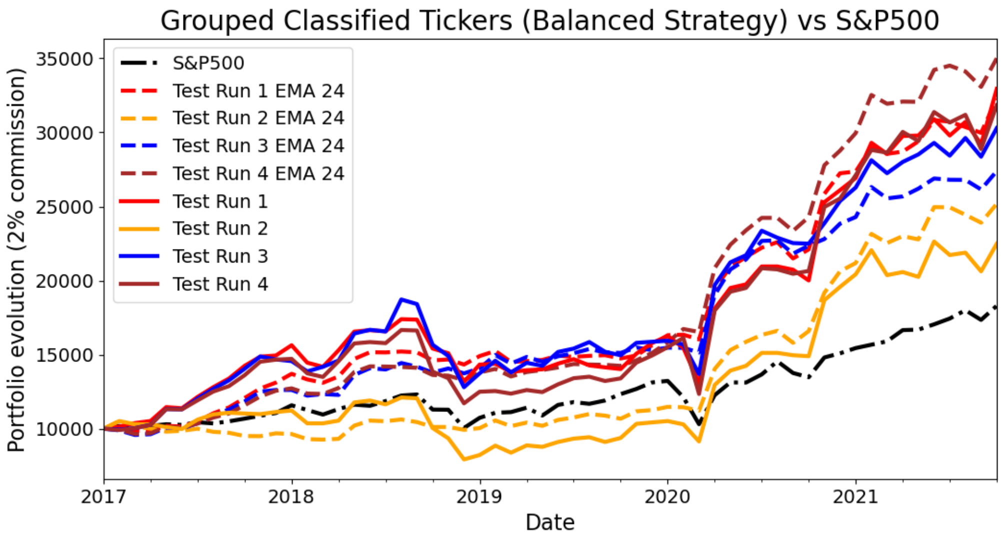

Figure 4.8: 7 Stock Portfolio evolution (Balanced) vs the S&P500 (y-axis - €)

best and worst Test Runs dipped in performance with the gap between them being reduced to 84.3 percentage points. Overall, based on these Test Runs there doesn't appear to exist any upside to picking this portfolio over any of the other portfolios. Depending on the Test Run the impact of the EMA trading system was either positive or negative. Figure 4.11 plots the various Test Runs for the **Balanced Strategy**.

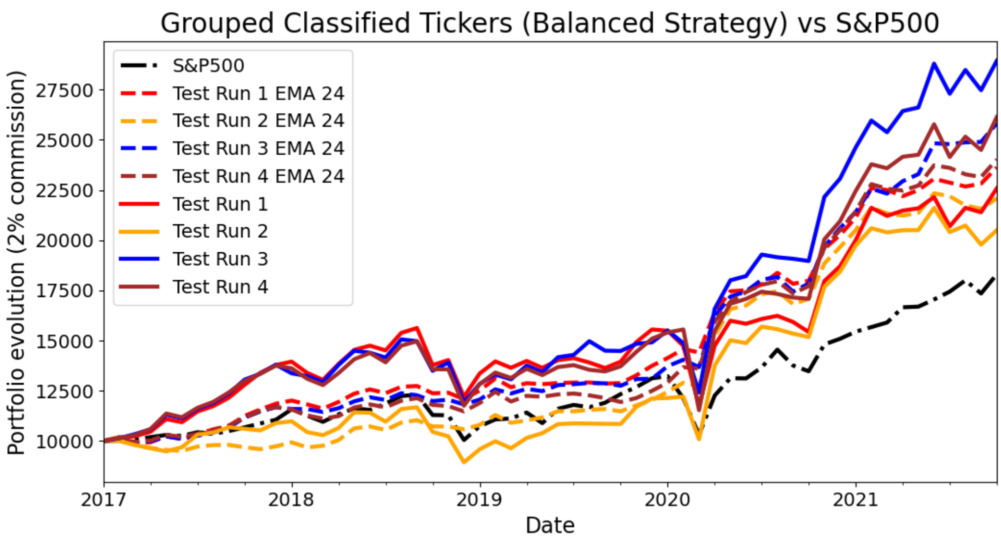

Figure 4.11: Sectors Stock Portfolio evolution vs the S&P500 (y-axis - €)

For the **Value Strategy** the average compound ROI was 188.7%, the lowest of all portfolios for this strategy. The gap between best and worst Test Run was also the smallest, being 25.6 percentage points,

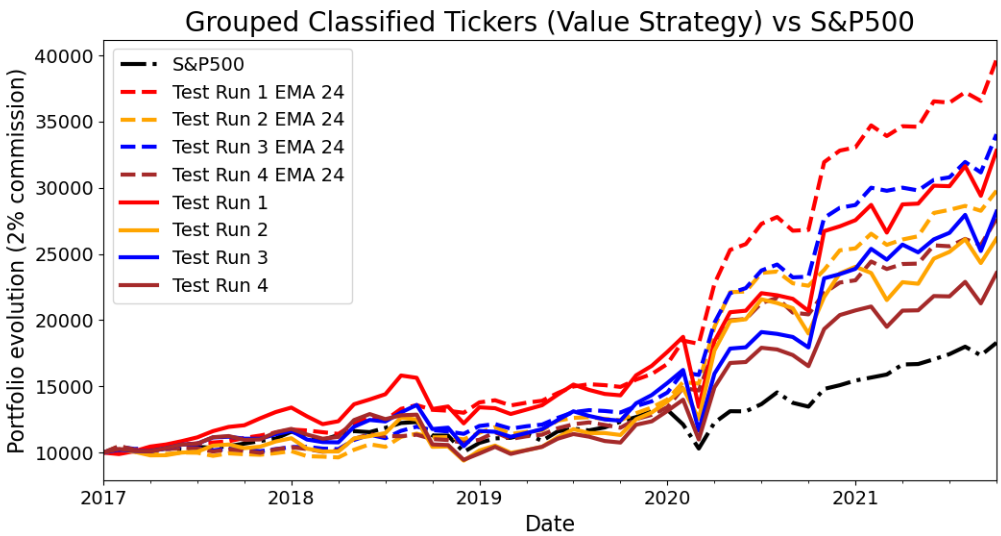

Figure 4.9: 7 Stock Portfolio evolution (Value) vs the S&P500 (y-axis - €)

indicating that this portfolio is extremely stable. Based on these Test Runs this portfolio appears to still be a viable portfolio that will yield good returns, yet it lacks the potential that other portfolios analysed provide. Figure 4.12 plots the various Test Runs for the **Value Strategy**.

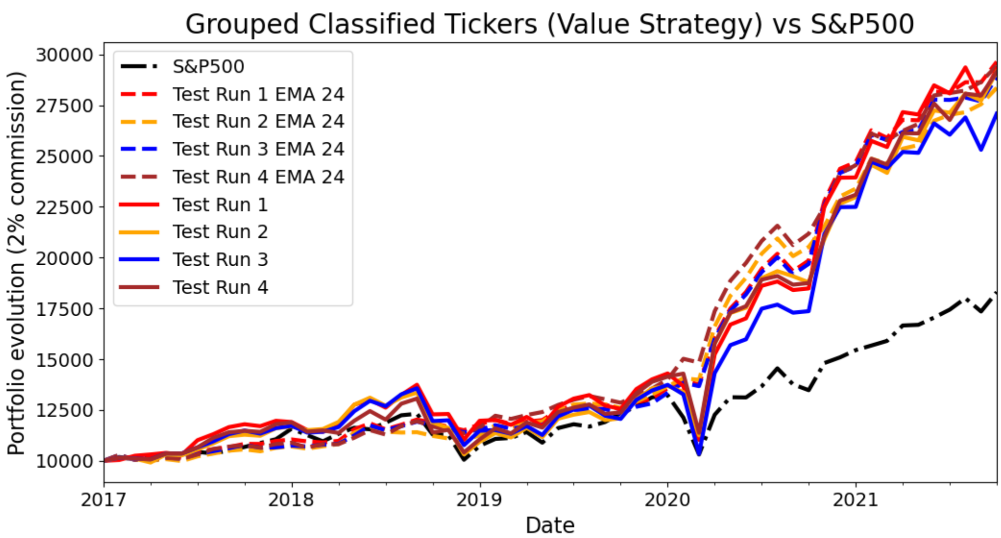

Figure 4.12: Sector Stock Portfolio evolution vs the S&P500 (y-axis - €)

Of all three strategies, for the SSP, the best strategy was the **Value Strategy**, followed by the **Balanced Strategy** and finally the **Growth Strategy**, a trend that remained consistent over all portfolio types. Overall this portfolio type doesn't appear to benefit the investor in any particular way barring maybe the increased stability of the **Growth Strategy**.

Table 4.7: BBP, PMP, number of trades and SR for strategies for the 7SP

Strategy	Test Run	BBP	PMP	Number of trades	SR
Growth Strategy	1	43.860%	57.895%	313	0.65
	1 EMA	54.386%	66.667%	1004	0.65
	2	38.596%	61.404%	301	0.61
	2 EMA	43.860%	63.158%	980	0.61
	3	50.877%	70.175%	317	0.61
	3 EMA	47.368%	66.667%	1003	0.61
	4	56.140%	64.912%	283	0.63
	4 EMA	56.140%	66.667%	999	0.63
Balanced Strategy	1	54.386%	66.667%	233	0.60
	1 EMA	56.140%	66.667%	993	0.60
	2	49.123%	59.649%	219	0.60
	2 EMA	47.368%	52.634%	1015	0.60
	3	61.404%	61.404%	217	0.60
	3 EMA	52.632%	63.158%	978	0.60
	4	57.895%	66.667%	251	0.60
	4 EMA	50.877%	61.404%	992	0.60
Value Strategy	1	59.649%	68.421%	241	0.69
	1 EMA	57.895%	70.175%	936	0.69
	2	50.877%	64.912%	231	0.68
	2 EMA	52.632%	63.158%	999	0.68
	3	49.123%	61.404%	241	0.68
	3 EMA	54.386%	66.667%	973	0.68
	4	43.860%	64.912%	239	0.68
	4 EMA	52.632%	63.158%	1002	0.68

As figure 4.10 indicates, all Test Runs performed similarly to the benchmark throughout the years for the **Growth Strategy**, with the Test Runs not applying the trading system performing better before the 2018 market decline and the Test Runs taking advantage of the EMA system mitigating the crash of 2020. The **Balanced Strategy** presented a mostly better performance than the benchmark, especially after the 2020 market crash as per figure 4.11. It was also able to find decent growth before the 2018 market dip, mainly for the Test Runs that didn't apply the trading system. Up until the 2020 market crash all **Balanced Strategy** Test Runs mostly acted as a consistent market tracker, with results never deviating much from the benchmark. As figure 4.12 shows, an explosion in returns also occurred in 2020, which was consistent among Test Runs.

The SSP proved to be the portfolio type that had the worst SR results, with the best and worst strategies being the same.

Table 4.8: ROI for all strategies for the SSP including total compound returns

Strategy	Test Run	2017	2018	2019	2020	2021	Total
Growth Strategy	1	20.4%	-5.3%	13.1%	16.9%	24.7%	88.2%
	1 EMA	14.8%	2.4%	9.8%	22.8%	25.3%	98.7%
	2	32.4%	-10.4%	14.5%	15.7 %	25.1%	96.7%
	2 EMA	18.9%	-0.1%	7.6%	20.7%	21.1%	86.9%
	3	27.3%	-11.2%	18.7%	6.5%	32.9%	89.9 %
	3 EMA	16.3%	0.1%	8.2%	14.7%	23.4%	78.3%
	4	19.6%	-3.2%	14.7%	19.0%	24.5%	96.9%
	4 EMA	15.7%	4.6%	10.1%	20.6%	22.2%	96.3%
Balanced Strategy	1	39.5%	-4.2%	15.9%	29.1%	12.9%	125.8%
	1 EMA	20.2%	3.8%	12.9%	49.5%	11.8%	135.2%
	2	9.8%	-12.7%	26.6%	62.5%	3.8%	105.0%
	2 EMA	-0.5%	8.7%	15.3%	64.6%	7.7%	121.1%
	3	33.72%	-3.94%	20.7%	58.7%	17.5%	189.3%
	3 EMA	16.0%	4.0%	13.8%	55.9%	20.6%	158.1%
	4	36.7%	-5.8%	20.1%	45.7%	16.4%	161.2%
	4 EMA	15.7%	2.8%	8.6%	65.9%	12.1%	140.1%
Value Strategy	1	19.1%	0.5%	19.5%	67.5%	22.2%	192.7%
	1 EMA	10.5%	6.7%	15.1%	81.8%	20.0%	196.4%
	2	16.9%	-6.5%	25.1%	67.9%	26.9%	191.5%
	2 EMA	7.1%	5.2%	21.4%	71.2%	21.2%	183.5%
	3	17.2%	-2.1%	19.7%	63.7%	20.4%	170.8%
	3 EMA	7.5%	7.3%	15.7%	83.5%	17.9%	188.7%
	4	18.9%	-7.0%	27.9%	63.3%	26.2%	191.4%
	4 EMA	8.9%	6.8%	20.4%	75.3%	19.9%	194.4%
S&P 500	-	15.8%	-7.3%	23.2%	16.8%	18.5%	83.0%

4.7 General Conclusions

First and foremost it can be confidently said that, upon analysing the results obtained throughout the various experiments, the **Value Strategy** was the best strategy employed, followed up by the **Balanced Strategy** and then the **Growth Strategy**, this over the four years of testing. This statement is mostly true when considering the period after **2020**, displaying the strength of picking value based stocks during market downtrends and subsequent rises. Although generally inferior, the **Balanced Strategy** still proved to be able to be a very sound strategy, capable of yielding great results, especially for years like **2017** and **2020**. Despite finding some difficulties when faced with market downtrends, this strategy proved to be generally capable of recovering during the following rises. Finally, the **Growth Strategy** proved to be a very unstable strategy, with possibilities of very high highs and extremely low lows, making it the less ideal strategy. Although it presented extremely good results during stable years such as **2017**, it suffered greatly during market downtrends and recoveries like in **2018** and **2020**, which could be expected since growth companies are usually more volatile and less capable of withstanding big blows.

Table 4.9: BBP, PMP, number of trades and SR for strategies for the SSP

Strategy	Test Run	BBP	PMP	Number of trades	SR
Growth Strategy	1	49.123%	61.404%	441	0.61
	1 EMA	50.877%	63.158%	1602	0.61
	2	52.632%	66.667%	409	0.59
	2 EMA	50.877%	64.912%	1596	0.59
	3	47.368%	64.912%	431	0.60
	3 EMA	43.860%	66.667%	1634	0.60
	4	42.105%	57.895%	387	0.60
	4 EMA	45.614%	66.667%	1566	0.60
Balanced Strategy	1	59.649%	66.667%	323	0.58
	1 EMA	54.386%	71.930%	1562	0.58
	2	50.877%	57.895%	349	0.56
	2 EMA	50.877%	63.158%	1594	0.56
	3	63.158%	63.158%	329	0.58
	3 EMA	52.631%	70.175%	1557	0.58
	4	66.667%	66.667%	317	0.57
	4 EMA	50.877%	63.158%	1610	0.57
Value Strategy	1	61.404%	66.667%	347	0.69
	1 EMA	49.123%	70.175%	1504	0.69
	2	59.649%	64.912%	291	0.68
	2 EMA	45.614%	71.930%	1494	0.68
	3	61.404%	66.667%	329	0.68
	3 EMA	50.877%	71.930%	1514	0.68
	4	63.158%	68.421%	307	0.68
	4 EMA	50.877%	71.930%	1485	0.68

Regarding the portfolio types considered, these must be analysed through the strategy's lens. For the **Growth Strategy**, the 2SP appears to be the portfolio capable of providing the best and worst possible results, so investing in such a portfolio using this strategy seems to be an extraordinarily risky endeavour. As the number of stocks increases the results seem to generally stabilize, culminating in the SSP that almost mimics the S&P 500 for all Test Runs. The rise in stability isn't always a positive sign, as can be observed for the **Balanced Strategy**, where the increase in stocks causes a clear dip in average returns without significantly improving the worst Test Runs, seemingly not being advantageous for the investor. For this strategy, based on the tests conducted, investing in a 2SP or 7SP appears to be the best advice. A similar and perhaps exacerbated trend can also be found for the **Value Strategy**, with not only the average returns decaying but also a decrease in returns for the worst Test Run with the increase of stocks, barring the SSP. For this strategy the clear best procedure is to invest on a 2SP or 5SP.

When it comes to the EMA trading system, the decision to implement it is mostly dependent on the type of risk the investor is willing to take. This system showed that it can be incredibly efficient, especially during market downtrends, improving negative results significantly with the other side of the coin being that it generally reduced the possible returns of really positive results. Thus, if the investor is looking for less volatile results, even if at the cost of better returns, the implementation of the EMA trading system is crucial.

5

Conclusions

Contents

This chapter will provide final conclusions regarding the work done during this thesis, focusing on results, system limitations and future work suggestions to further improve the model.

5.1 Conclusions

In this thesis a GA based stock selection model paired with a trading system was created in order to take advantage of patterns found in the market. Three main strategies were analysed, a **Growth Strategy** that mainly focuses on the potential of companies, a **Balanced Strategy** that will take into consideration all kinds of companies and a **Value Strategy** that searches for companies that have strong and stable positions in the market. The companies considered were all inserted in the S&P 500 and overall the training and testing period ranged from 31-01-2017 up to 31-10-2021. The portfolios considered for testing, applied to each strategy, were the 2SP, the 5SP, the 7SP and the SSP, with the stocks either being equally distributed and updated every month or equally distributed but traded utilizing an EMA based trading system, although monthly stock updates were still applied. The results were overall extremely positive with most strategies being able to find great results when compared to the benchmark. Given that the portfolios generated are fairly simple these results were mainly due to the GA based stock selection model, proving its capabilities of finding good stocks. It is also important to highlight the impact of the EMA trading system, specially for the **Growth Strategy**, which was able to severely diminish losses during more volatile months.

Overall this thesis was able to accomplish several objectives. Firstly data regarding S&P 500 companies' financial statements was successfully acquired and utilized to calculate various FIs for multiple months. Then, utilizing a K-means clustering algorithm the performance of a company's FIs was classified for every end of the month in the data range, this done by contrasting with sector peers. Following this step a GA was taken advantage of to search for the importance that each indicator has in a company's performance, utilizing a previous three month price variation method as the fitness function. All these steps culminated in the creation of a stock selector that takes into consideration both the performance of FIs among industry peers and the importance that each fundamental indicator must have. This classifier was updated at the end of every month. Finally the stock selector was tested by forming various different portfolios and analysing the results for various strategies, even introducing a trading system based on the EMA.

5.2 Future Work

Although the model created during this thesis yielded good returns there is always room for improvement. Regarding future work, these are some suggestions that are believed to further improve the model developed:

- Test the model with different types of evolutionary algorithms;
- Explore different trading systems with the assistance of more TIs;
- Apply different types of portfolio techniques like the Modern Portfolio Theory [24] in order to better optimize portfolio weights;
- Explore the application of the GA Module developed in this thesis to find the worst companies instead of the best companies, opening shorting options;
- Take into account different types of trading costs in order to get a better grasp of the overall performance;
- Explore different markets besides the S&P 500;
- Apply different methods of clustering, perhaps hierarchical, for the classification.

Bibliography

[1] B. G. Malkiel, "The efficient market hypothesis and its critics," *Journal of Economic Perspectives*, vol. 17, pp. 59–82, 2003.

[2] K. Hrazdil, K. Trottier, and R. Zhang, "A comparison of industry classification schemes: A large sample study," *Economics Letters*, vol. 118, 2012.

[3] R. C. Cavalcante, R. C. Brasileiro, V. L.F. Souza, J. P. Nobrega, and A. L.I. Oliveira, "Computational intelligence and financial markets: A survey and future directions," *Expert Systems With Applications*, vol. 55, pp. 194–211, 2016.

[4] W. Jiang, "Applications of deep learning in stock market prediction: Recent progress," *Expert Systems With Applications*, vol. 184, 2021.

[5] Y. Hu, K. Liu, X. Zhang, L. Su, E. Ngai, and M. Liu, "Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review," *Applied Soft Computing*, vol. 36, pp. 534–551, 2015.

[6] E. Beyaz, F. Tekiner, X. Zeng, and J. A. Keane, "Comparing technical and fundamental indicators in stock price forecasting," 2018.

[7] A. Slowik and H. Kwasnicka, "Evolutionary algorithms and their applications to engineering problems," *Neural Computing and Applications*, vol. 32, 2020.

[8] B. Graham and D. L. Dodd, *Security Analysis*. McGraw-Hill, 1934.

[9] B. Graham, *The Intelligent Investor*. Harper & Brothers, 1949.

[10] M. Buffett and D. Clark, *Warren Buffett and the Interpretation of Financial Statements: The Search for the Company with a Durable Competitive Advantage*. Scribner, 2008.

[11] S.-S. Chen, "Predicting the bear stock market: Macroeconomic variables as leading indicators," *Journal of Banking & Finance*, vol. 33, pp. 211–223, 2009.

[12] S. B. Achelis, *Technical Analysis from A to Z*. McGraw-Hill, 2001.

[13] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, p. 651–666, 2010.

[14] D. Wu, X. Wang, and S. Wu, "Construction of stock portfolios based on k-means clustering of continuous trend features," *Knowledge-Based Systems*, vol. 252, 2022.

[15] J. Sass and A.-K. Thös, "Risk reduction and portfolio optimization using clustering methods," *Econometrics and Statistics*, 2021.

[16] M. Khedmati and P. Azin, "An online portfolio selection algorithm using clustering approaches and considering transaction costs," *Expert Systems with Applications*, vol. 159, 2020.

[17] Y.-N. Lu, S.-P. Li, L.-X. Zhong, X.-F. Jiang, and F. Ren, "A clustering-based portfolio strategy incorporating momentum effect and market trend prediction," *Chaos, Solitons and Fractals*, vol. 117, pp. 1–15, 2018.

[18] C. C. Aggarwal and C. K. Reddy, *Data Clustering Algorithms and Applications*. Chapman & Hall, 2014.

[19] L. Rutkowski, *Computational Intelligence: Methods and Techniques*. Springer, 2008.

[20] S. Lim, M. Kim, and C. W. Ahn, "A genetic algorithm (ga) approach to the portfolio design based on market movements and asset valuations," *IEEE Access*, vol. 8, 2020.

[21] A. Silva, R. Neves, and N. Horta, "A hybrid approach to portfolio composition based on fundamental and technical indicators," *Expert Systems with Applications*, vol. 42, 2015.

[22] D. Cheong, Y. M. Kim, H. W. Byun, K. J. Oh, and T. Y. Kim, "Using genetic algorithm to support clustering-based portfolio optimization by investor information," *Applied Soft Computing*, vol. 61, 2017.

[23] B. C. Barroso, R. Cardoso, and M. K. Melo, "Performance analysis of the integration between portfolio optimization and technical analysis strategies in the brazilian stock market," *Expert Systems With Applications*, vol. 186, 2021.

[24] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, 1952.

[25] C.-F. Huang, "A hybrid stock selection model using genetic algorithms and support vector regression," *Applied Soft Computing*, vol. 12, 2012.

[26] X. Meng and X. Chen, "The investment models based on the fundamental and technical indicators of listed companies," *37th Chinese Control Conference*, 2018.

[27] T. R. C. C. da Costa, R. T. Nazário, G. S. Z. Bergo, V. A. Sobreiro, and H. Kimura, “Trading system based on the use of technical analysis: A computational experiment,” *Journal of Behavioral and Experimental Finance*, vol. 6, pp. 42–55, 2015.

[28] “Macrotrends,” https://www.macrotrends.net/.

[29] “yfinance library,” https://pypi.org/project/yfinance/.

[30] “scikit-learn - machine learning in python,” https://scikit-learn.org/stable/.

A

Appendix A - List of Chromosomes

The Test Run chromosomes obtained for this work will be presented with the values being rounded up to two decimal places.

A.1 Growth Strategy

For the **Growth Strategy** the weights w_i refer to the importance that should be given to the following indicators: w_1 - D/E; w_2 - ROE; w_3 - NPM; w_4 - RG; w_5 - RR.

Test Run	w_1	w_2	w_3	w_4	w_5	**Fitness**
Test Run 1	-4.74	3.70	1.64	4.57	2.92	36.32
Test Run 2	-3.84	2.35	0.16	3.17	4.85	36.83
Test Run 3	-4.85	3.10	1.00	3.75	2.58	37.12
Test Run 4	-2.74	4.23	1.28	4.20	3.38	33.38

Table A.1: Growth Strategy Test Run Chromosomes (2017)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	-0.57	1.69	0.30	2.26	2.65	33.84
Test Run 2	-1.58	2.99	0.81	3.56	0.67	36.90
Test Run 3	-2.16	4.10	0.93	4.62	0.98	37.93
Test Run 4	-3.02	2.74	1.39	4.83	0.76	34.58

Table A.2: Growth Strategy Test Run Chromosomes (2018)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	-0.81	1.34	0.04	3.23	3.74	38.36
Test Run 2	-3.33	3.37	-2.85	1.25	4.92	38.39
Test Run 3	-1.03	2.29	-0.39	4.66	4.81	39.14
Test Run 4	-1.57	2.18	-0.68	3.90	4.62	36.88

Table A.3: Growth Strategy Test Run Chromosomes (2019)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	-0.65	2.39	-2.27	4.74	-0.75	34.92
Test Run 2	-0.31	1.03	-1.01	3.18	0.22	35.61
Test Run 3	-0.96	3.41	-2.05	4.49	-1.52	32.98
Test Run 4	-0.60	2.47	-2.19	4.82	-0.62	33.22

Table A.4: Growth Strategy Test Run Chromosomes (2020)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	-1.64	2.16	0.92	3.33	-0.98	31.33
Test Run 2	-0.01	0.24	-0.23	2.49	0.12	38.17
Test Run 3	0.99	0.07	2.58	4.50	-1.08	31.66
Test Run 4	-1.32	2.18	-1.26	4.92	-0.30	31.11

Table A.5: Growth Strategy Test Run Chromosomes (2021)

A.2 Balanced Strategy

For the **Balanced Strategy** the weights w_i refer to the importance that should be given to the following indicators: w_1 - D/E; w_2 - CR; w_3 - EPS; w_4 - ROE; w_5 - ROA; w_6 - NPM; w_7 - GPM; w_8 - RG; w_9 - OM.

Test Run	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	Fitness
Test Run 1	-3.73	1.02	-1.20	3.07	1.43	1.76	0.60	4.79	-2.68	39.94
Test Run 2	3.22	-0.67	-4.09	2.45	-2.81	-2.99	-3.02	4.60	1.42	33.55
Test Run 3	-3.68	1.37	-1.12	4.04	0.49	3.59	0.27	4.13	-3.30	40.32
Test Run 4	-4.02	1.97	-1.07	4.99	1.76	4.04	1.41	4.59	-4.90	39.60

Table A.6: Balanced Strategy Test Run Chromosomes (2017)

Test Run	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	Fitness
Test Run 1	-3.33	0.13	-1.04	0.93	1.32	2.59	-0.16	4.54	-1.84	36.98
Test Run 2	-3.19	0.34	-3.18	4.40	1.75	1.29	-0.26	4.34	-1.05	39.09
Test Run 3	-3.55	1.39	-1.76	3.75	1.06	1.72	-0.14	4.74	-2.63	41.49
Test Run 4	-3.06	0.34	-2.01	2.39	1.54	2.67	0.43	4.67	-2.32	38.31

Table A.7: Balanced Strategy Test Run Chromosomes (2018)

Test Run	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	Fitness
Test Run 1	-2.68	1.88	-2.14	4.86	0.17	1.53	1.43	4.99	-2.25	43.67
Test Run 2	0.14	2.71	-1.23	1.59	-0.69	1.32	2.13	4.88	-1.32	44.14
Test Run 3	-0.97	2.52	-3.47	1.99	-1.07	1.25	3.12	4.52	-1.50	43.19
Test Run 4	-2.39	2.12	-3.16	4.19	2.27	-0.35	3.32	4.71	-2.58	43.68

Table A.8: Balanced Strategy Test Run Chromosomes (2019)

Test Run	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	Fitness
Test Run 1	0.52	1.98	-4.24	2.40	-0.13	0.83	2.23	4.81	-1.22	42.54
Test Run 2	-0.12	3.00	-3.79	3.72	-1.03	0.61	2.45	4.85	-1.47	43.20
Test Run 3	-0.88	3.56	-2.44	3.42	-2.70	2.35	1.00	4.93	-1.22	40.08
Test Run 4	0.01	2.23	-3.24	3.46	-1.82	2.31	1.26	4.56	-2.00	42.45

Table A.9: Balanced Strategy Test Run Chromosomes (2020)

Test Run	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	Fitness
Test Run 1	-0.52	2.38	-4.72	2.00	1.34	1.21	-0.89	3.91	-1.09	41.23
Test Run 2	0.96	2.36	-3.28	1.35	0.67	2.19	-1.00	4.91	-2.50	41.47
Test Run 3	0.28	3.44	-2.50	3.30	-1.07	0.96	1.99	4.33	-1.96	39.76
Test Run 4	0.80	1.96	-3.72	1.58	1.45	0.19	-0.41	4.27	-0.09	40.62

Table A.10: Balanced Strategy Test Run Chromosomes (2021)

A.3 Value Strategy

For the **Value Strategy** the weights w_i refer to the importance that should be given to the following indicators: w_1 - CR; w_2 - EPS; w_3 - ROA; w_4 - RG; w_5 - PR.

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	-0.32	-3.42	1.90	3.74	-4.80	35.67
Test Run 2	-0.55	-3.00	-0.52	4.09	-4.97	36.80
Test Run 3	-0.88	-3.53	0.01	4.35	-4.84	41.64
Test Run 4	-0.48	-1.94	-0.43	2.80	-3.64	37.06

Table A.11: Value Strategy Test Run Chromosomes (2017)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	-1.56	-3.11	3.17	4.86	-3.05	34.18
Test Run 2	-1.30	-3.61	3.73	4.97	-1.64	37.33
Test Run 3	-1.83	-2.75	3.66	4.63	-3.12	34.89
Test Run 4	0.01	-0.94	3.52	4.53	-1.54	35.40

Table A.12: Value Strategy Test Run Chromosomes (2018)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	1.78	-2.67	1.07	4.62	-3.91	43.38
Test Run 2	1.52	-2.56	1.37	3.51	-3.16	43.74
Test Run 3	1.37	-2.04	0.80	3.54	-2.87	43.14
Test Run 4	1.70	-2.25	1.28	4.91	-2.56	42.87

Table A.13: Value Strategy Test Run Chromosomes (2019)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	1.22	-2.60	0.99	3.06	-1.90	40.54
Test Run 2	3.18	-1.75	1.07	4.35	-3.44	38.37
Test Run 3	0.75	-1.50	0.55	1.74	-1.06	40.49
Test Run 4	1.40	-3.32	1.57	3.32	-2.37	39.96

Table A.14: Value Strategy Test Run Chromosomes (2020)

Test Run	w_1	w_2	w_3	w_4	w_5	Fitness
Test Run 1	0.75	-1.74	1.17	3.16	-3.18	44.49
Test Run 2	0.43	-4.72	4.11	3.71	-4.39	44.35
Test Run 3	1.55	-4.05	1.91	4.01	-3.02	43.36
Test Run 4	0.17	-3.27	2.22	3.82	-4.29	43.66

Table A.15: Value Strategy Test Run Chromosomes (2021)