Handelshøgskolen ved UiT

A comparison of strategies for portfolio allocation

Could advanced statistical methods for portfolio construction generate excess return in the stock market?

Gabriel Karlsen og Sondre Aarberg Wara

Masteroppgave i Økonomi og Administrasjon, BED-3901, mai 2022

Abstract

This study compares five different strategies for asset allocation using five different ETFs from the US stock market. Two strategies, Buy and Hold and the Naive 1/N portfolio, do not consider any change in the dataset and therefore is the easiest two strategies. One portfolio is constructed using the Modern Portfolio Theory (MPT) and is the optimal portfolio for minimum variance. The other portfolio from MPT is the tangency portfolio (also called the Max Sharpe Ratio Portfolio). The last strategy is the most complex one and is constructed using the Random Forest theory within Machine Learning. All portfolios, except the buy and hold portfolio, are rebalanced monthly (72 times), with a rolling window of data from the previous 12 months. In total we look at six years (2016-2021) of daily returns for each ETF to calculate our allocations and all calculations were done in either excel or R.

The first four strategies show a similar return between 16,4% to 17,8% annually while the Random Forest portfolio yields an annual return of 24,6% during the 6 years we study. Interestingly, the Random Forest portfolio is also the least risky portfolio with an annual standard deviation of 12,4% compared to the other portfolios with an annual volatility of about 15,5%. This implies that the Random Forest has a Sharpe-ratio of 1,99, the maximum Sharpe-Ratio portfolio got a Sharpe-Ratio of 1,15, whereas the other three portfolios have a Sharpe-Ratio of about 1,1. A regression analysis shows that for the first four strategies the Carhart 4 factor model can explain 95,8% - 97,9% of the variance in the portfolios while for the Random Forest the model can only explain 78,8%. Random Forest is also the only strategy with a statistically significant alpha (α) of almost 9% annually.

Our results from Random Forest as a method for portfolio allocation is in accordance with previous research and challenges both the weak and semi-weak market efficiency. It also shows that Machine Learning has a potential to detect patterns for excess return and should therefore be of interest to investors and traders.

Acknowledgments

This paper marks the end of five years of education at University of Tromsø, The Arctic University. The paper withholds 30 ECTS credits and is written within our master's minor field, Finance. Although, this is not offered as a major subject in Tromsø, we appreciate the possibility to write our master's thesis within this subject.

For us it has been a long process of not only writing about the topics we have learned throughout our subjects at school, but also learn about one of the most complex and innovative subjects to date within the field of finance. After many weeks of frustrating attempts and googling about Machine Learning in R, we finally started to get some results and after coding the hardest part, the rest of the paper felt fairly quick and simple. We both feel the process itself have tough us a lot about coding, working strategically under time pressure and cooperating throughout the whole process.

This paper would not have been at all if it had not been for our teacher, Thomas Leirvik. His contribution with quick response and continuous follow-ups and have helped us immensely. His passion for the subject enthusiasm around our paper have kept us motivated and optimistic throughout the process. We would also like to thank Espen Sirnes, for his contribution to this thesis that have truly lifted the quality of this papers result and relevance. We would like to thank family and friends, not only for the duration of this paper, but for the whole study period of five years. We have been busy and stressed with seminars, assignments and exams and we would not have been able to complete this without your support and understanding throughout the years.

Lastly, we would like to thank the University of Tromsø and the teachers at Handelshøgskolen i Tromsø for 5 fantastic years.

Tromsø, May 29. 2022

Gabriel Kr

Gabriel Karlsen

Sondre Aa Wara

Sondre Aarberg Wara

Sammendrag på Norsk

Denne studien sammenligner fem forskjellige strategier for kapital allokering med fem forskjellige ETFer fra det amerikanske aksjemarkedet. To strategier, Buy and Hold og Naive 1/N portefølje, tar ikke endringer i datasettet til betraktning og er derfor de to enkleste strategiene. En portefølje er satt sammen ved bruk av Modern Portfolio Theory (MPT) og er den optimale porteføljen for minimal varians. Den andre porteføljen fra MPT er tangentporteføljen (også kalt maks Sharpe-Ratioporteføljen). Den siste strategien er den mest komplekse og er en Random Forest metode innen Machine Learning. Alle porteføljene, bortsett fra Buy and Hold, er rebalansert månedlig (72 ganger), med et rullerende vindu av data fra de siste 12 måneder. Totalt ser vi på de siste 6 årene (2016-2021) med daglige avkastninger for hver ETF til å regne kapitalallokering og all utregning er gjort i enten Excel eller R.

De første fire strategiene viser tilnærmet like avkastninger mellom 16,4-17,8% årlig mens Random Forest porteføljen gir en total avkastning på 24,6% årlig. Random Forest er også den porteføljen med lavest risiko med et standardavvik på 12,4% årlig mot de andre porteføljene med 15,5% årlig, noe som gir Random Forest en Sharpe-Ratio på 1,99 og de øvrige porteføljene en Sharpe-Ratio mellom 1,07-1,09. Sharpe-Ratioen for maks Sharpe-Ratio porteføljen ble 1,15. En regresjonsanalyse viser at for de første fire strategiene kan Fama French Carharts 4-faktor modell forklare 95,8%-97,9% av variansen til, mens for Random Fortest kan modellen bare forklare 78,8%. Random Forest er også den eneste strategien med en statistisk signifikant alfa (α) på nesten 9% årlig.

Våre resultater fra Random Forest som en metode for porteføljesammensetting er i samsvar med tidligere forskning og utfordrer både svak og semi-svak markedseffisiens. Den viser også at Machine Learning har et potensial til å se mønstre for meravkastning og burde derfor være av interesse for investorer og tradere.

Nøkkelord

Market efficiency, Machine Learning, Random Forest, Markowitz Modern Portfolio Theory, Multiple Regression, Exchange Traded Funds

Table of Contents

Figures and Tables

Equations

1.0 Introduction

Portfolio selection have been a central question for investors since stocks were traded through papers on a trading floor. The advantages and possibilities of owning assets in more than one company are numerous and investors have been trying to exploit the advantages for many decades. The challenges to predicting future stock prices, the company's risk and the markets overall risk have proven to be difficult. The stock market seems divided between those who claim asset allocation in a portfolio is the most important factor to portfolio returns (Brinson et al., 1986; Shukla & Bogle, 1994) and those who claim it is better to focus on finding a few stocks that can outperform the market to consistently beat the markets median return (Tan et al., 2019).

To find a consistent and reliable explanation to stock returns and predict future movements was the driving question for finance academics and led to classical models like CAPM, Fama French 3-factor and Carhart 4-factor model that claimed portfolio returns could be explained by factors such as market risk, market capital and momentum (Carhart, 1997; Fama & French, 1993; Sharpe, 1964). These models have a similarity with their linearity between return and their explaining factors. This fact has been a major factor for the following academic works within finance. This linearity is however a simplification of the complexity that is the stock market, and the ability to predict future movement in the market is therefore difficult on the basis of these factors (Zhu et al., 2012).

For this reason, new models that challenged the linearity approach was necessary and machine learning was introduced. Innovation within technology has led to the discovery of new and complex models within finance, one of them being heavy statistical models such as machine learning. These methods enable simulation and modeling of problems which have been too complex to solve previously. Through powerful models such as Artificial Neural Networks (Guresen et al., 2011; Qiu et al., 2016), Support Vector Machines (Huang et al 2005; Tay & Cao 2001) and Random Forrest (Krauss et al 2017), the ability to explain and predict future returns have significantly improved. The development in machine learning methods have let to alternative methods for illustrating the relationship between return and their relevant factors, and thus let to a model-diversification compared to Modern Portfolio Theory (MPT).

The usage of machine learning within finance are not yet a broad concept to the academic environment and there are few written papers on the topic compared to other aspects well known in financial literature. There are however a lot of published articles related to programming, statistics, and data science (Henrique et al., 2019)

Even though there are a large volume of articles predicting future stock prices with machine learning, few of them illustrate the relationship between the stock selection and their underlying factors with machine learning. A search for literature online reveals that out of twenty machine learning articles, only one of them is in the field of economy/finance (Henrique et al., 2019).

Even with this new technology to help predict stock movement, some investors claim long term buy and hold strategies are favorable to the advanced strategies with optimalization and relatively frequent rebalancing (Qian, 2014). Much due to the risk of mistiming the market and even more because of the cost with rebalancing. In fact, DeMiguel et al. (2009) show that the naive 1/N portfolio is equal or better performing than the rebalancing portfolios using modern portfolio theory, or Markowitz weights.

In 2015 Hillard & Hillard found that no matter what monetary policy is at the time, either with rising or decreasing interest rate, the buy and hold strategy did not beat portfolios optimized with Markowitz weights. This contrasted with DeMiguel (2009) and could be due to a number of factors. It shows however that results regarding portfolio allocation is dependent on what data you chose and what method you use to analyze the data. With the introduction of machine learning within finance the question still stands.

1.1 Problem statement

In our introduction we explained how two simple strategies for asset allocation could potentially yield the same return as optimizing strategies that include more sophisticated calculation and simulation. The two studies we mentioned (DeMiguel and Hillard & Hillard) found results that were in contrast with one another, and it enlightens the challenges portfolio managers face when choosing a strategy for their investors. With the innovation and implementation of machine learning many investors sees the potential for arbitrage opportunities, while some claim the markets are too efficient to spend time and money on complex models and calculations. The research question will therefore be formulated as such:

Can advanced statistical methods for portfolio construction generate excess return in the stock market?

We also include the question:

1. Do the strategies yield an excess return, or could it be explained by Carhart's four-factor model?

To look at this problem statement we start by breaking down previous research relevant for this thesis and explain what angle we will take for this paper.

1.2 Previous Research

To map the previous findings, we have chosen four papers based on their theory and methodology. Our paper will be a combination of all of these. For the naive 1/N strategy we have used the first paper from DeMiguel et al. (2009). Markowitz theory and the efficient frontier portfolios we have used the article from Hillard & Hillard (2015). We will also use the results from both previous studies and compare our findings with these. For the Random Forest theory, methodology and findings we have chosen Tan et al. (2019). We include Kilskar (2019) to compare results.

Table 1: Previous Studies and Their Findings

Authors	Title	Findings
DeMiguel et al. (2009)	Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?	The naive 1/N policy to portfolio allocation is equally good or better than optimizing portfolios. At least for the portfolios rebalanced based on Markowitz portfolio theory.
Hillard & Hillard (2015)	A comparison of Rebalanced and Buy and Hold Portfolios: Does Monetary Policy Matter?	Rebalancing strategies beats buy and hold and holds for all measures.
Tan et al. (2019)	Stock selection with Random Forest: An exploitation of excess return in the Chinese stock market	Machine learning is not yet a big field within economics and finance. The test of fundamental/technical factors and momentum factors points to a less

		efficient market and promising return. Although the return seems to be decreasing for every year, it still suggests for potential excess return and a strong strategy for both lower risk and high return.
Kilskar (2019)	Aksjeutvelgelse ved bruk av Random Forest: En maskinlæringstilnærming til meravkastning	The Random Forest strategy still seems strong for technical factors and momentum and although a good return the fundamental factors seems the weakest out of all the tested factors for predicting future returns.

From DeMiguel et al. we find that the naive 1/N weighted portfolio is equally if not better performing than the optimal weighted portfolio, under the Markowitz theory. Hillard & Hillard on the other hand finds that the optimized portfolios beat the naive 1/N and Buy and Hold on all measures. They mention some reasons for this might be sampling error and the mixture of stocks and bonds chosen. Tan et al. finds promising results with excess return in the Chinese market and Kilskar's thesis indicates the same, but shows poor results for fundamental analysis factors. Therefore, we have a good foundation to investigate an innovative and somewhat inconsistent subject between academics, of what capital allocation strategy is better for the performance of a portfolio.

Our study will be a hybrid of these four previous studies. We take strategies from the first three articles and combine them in to one study. We compare our findings and in addition to compare results with the market (S&P 500) we also do a regression analysis to see how much of our return can be explained by Fama French Carhart's 4-factors and how much, if any, is actual excess return from the strategy.

2 Theory

In this chapter we will present some of the fundamental theory needed to understand what we are looking for in our analysis and some of the concepts that will be used throughout this study.

2.1 Random Walk

Random walk is a stochastic process, that describes the independence of price movements. Previous prices cannot be used to determine future prices. Random Walk however indicated that today’s price is the best estimate of tomorrows price. The formula can be expressed as:

$$Y_t = Y_{t-1} + \varepsilon_t \qquad 2.1$$

Where Y is the price at time t, and ε_t is the white noise, which is a stationary process without autocorrelation. The initial value Y is independent white noise with t ≥ 1.

Fluctuations on prices on our data are triggered by new information in the market that changes expectations. In Random Walk theory this information is impossible to predict and thus, are random. If you could predict movements in the market based on previous trend patterns, it would suggest that the error term in the random walk equation is autocorrelated. Period t correlates with the movement to period t+1. (Bjordal & Opdahl, 2017) Figure 1 illustrates this further.

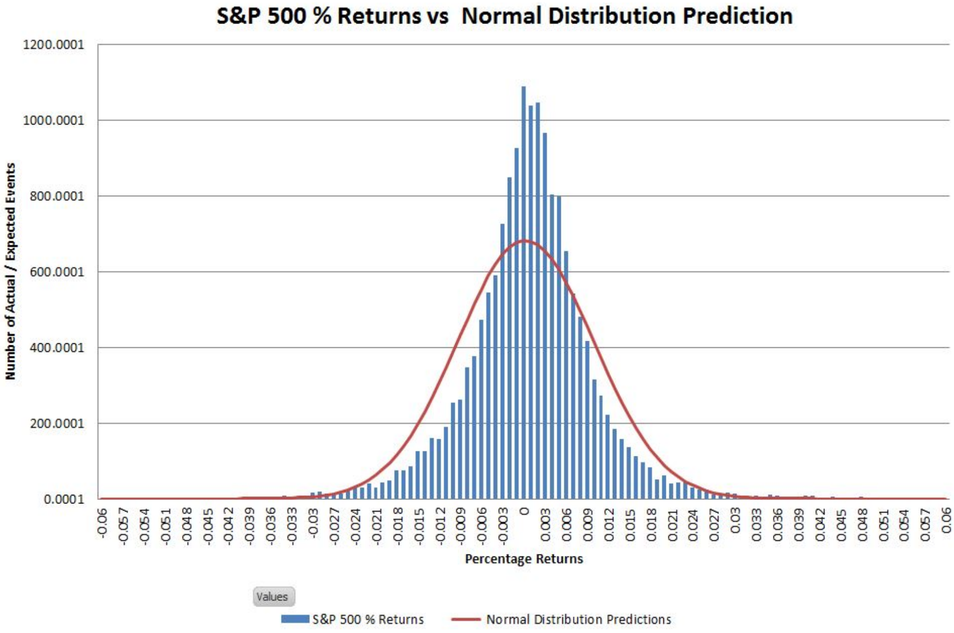

Figure 1: Normal Distribution Compared to the Empirical Distribution of the S&P 500

Figure 1 illustrates the normal distribution (red line) and the actual daily returns of the S&P 500 index last 50 years. From the figure we can see that most of the data lies around the mean or within ± 3 Standard Deviations. The prices are slightly right skewed witch implies a slight autocorrelation, between *t* and *t+1* as explained above. We can also see a few points where the tails breach the normal distribution line, +- 4 sigma, which means that low-probability in a normally distributed world, have a much more frequent occurrence in real-world observations, which means that the dataset we apply is slightly more fat-tailed than the normal distribution.

From this figure we can confirm that most observations are around the mean although low-probability events can have huge impact on the return. Although slightly skewed, the figure seems to confirm the random walk theory, and predictions on the data itself should be impossible. This fact however is not accepted by all investors and more than 90% claims to use technical indicators when making investment decisions (Utami & Nugroho, 2017).

2.2 Market Efficiency

To analyze our strategies return compared to the market we introduce theory about market efficiency. At the end of this paper, we will discuss how efficient the market really is when we look at our results.

Analyzing time series to identify patterns was one of the first uses of data analysis within finance. Through an analysis in attempt to find systematic correlation in the movements of price, Kendall and Hill (1953), found that movements in price seemed to be random and unpredictable as mentioned in previous chapter. This laid the foundation for theory on market efficiency, by establishing the understanding of a well-functioning, efficient market, (Bodie et al., 2018).

Today’s interpretation of the term Market efficiency where first introduced by Eugene F. Fama in (1970). He argued that the financial market holds all available information and that the prices reflect all information, historic prices and expectations. Thus, there is no excess return to be gained with technical analysis. Eugene Fama however acknowledges the term is misleading, and prefers a Market Efficiency Theorem, that can be challenged frequently (Menkhoff, 2010). This theory claims that investors cannot outperform the market and market anomalies should not exist because they will immediately be arbitraged away.

2.2.1 Grades of Efficiency

The market is usually divided into three forms of market efficiency: weak, semi-strong and strong efficiency.

Weak market efficiency implies that prices in the market reflects all available information. Historic prices, volume and interests for shorting (Bodie et al., 2018). This form of market efficiency implies that there is no arbitrage opportunity to be gained from technical analysis, or analysis to find patterns in market data, as it is already reflected in the market prices.

Semi-strong market efficiency is a further development of the weak efficient market. It adds on all publicly available information on the market prices. This includes company's product portfolio, leadership, balance sheet, patents, results prognoses and accounting practices (Eugene F. Fama, 1970). Under the semi-strong market efficiency, no arbitrage opportunity can be found using fundamental analysis, neither through income statements, balance sheets nor company news.

Under strong market efficiency the price of a stock reflects absolute all information about a company, this includes information only available to people inside the company. This form of market efficiency implies that there is no arbitrage opportunity possible to be found.
Famas theory assumes that for the strong efficient market, sum of all information is eliminated by transaction costs. The weaker form for efficiency, an arguably more realistic theory, claims that prices reflect information to the point where marginal advantages with information does not override the marginal costs (Fama, 1991).

2.3 Exchange Traded Funds

Exchange Traded Funds (EFTs) is one of the newest and most important innovations within finance in decades (Lettau & Madhavan, 2018). It is designed to track the performance of for example a stock market index, a segment of the market e.g. the clean energy sector, a region, e.g. the Euro area, a commodity, e.g. the gold price. It behaves much like a mutual fund does, but with some substantial differences to be explained below. It was first introduced in 1993 by State Street, the ETF was named SPDR, and was designed to track the S&P 500 index. It is still today the largest ETF with $ 358 billion in assets by May 2022 (Lettau & Madhavan, 2018). Today there are over 8,600 ETFs worldwide (May 2021). Later years ETFs consisting of options, futures, swaps and even cryptocurrencies have been blooming, although these are referred to as synthetic ETFs and will not be covered further in this paper (Norrestad, 2022).

2.3.1 Function and Differences

The reason for the innovation was that in an ETF, more investors now had access to a low cost means of gaining a diversified portfolio and the capacity for intraday trading. They also enabled regular investors to invest in a range of assets which may have otherwise been inaccessible or prohibitively expensive, such as gold, large blue-chip companies, and emerging markets.

ETFs are as mentioned above, similar to managed funds. However, investors can trade ETF shares frequently intraday on a stock exchange, while regular managed funds complete transactions, at most, once a day (Kosev & Williams, 2011). Because ETFs do not generally buy and sell the underlying assets to create shares, brokerage cost and management fees tend to be lower with ETFs. ETFs cannot be bought from the issuer itself but must be bought at a stock exchange. It is also worth mentioning that most mutual funds will have some room to outperform the set index, while ETFs will most of the time have the exact same exposure as the set index.

2.4 Statistical and Theoretical Concepts

To understand the concept behind the theory used in this thesis, we must know the theoretical, mathematical and statistical elements that these theories are based on. These are presented in the following under-chapters.

2.4.1 Returns

As an estimate of an assets return, we use the logarithm-returns. To calculate the daily logarithmic return, we take the logarithm of the price at time t divided by the price at time t-1 with daily observations. The formula is given as:

$$r_{i,d} = \ln\left(\frac{P_{i,t}}{P_{i.t-1}}\right). \qquad 2.2$$

Here $r_{i,\,d}$ is the daily logarithmic return for asset *i* at time *t*, $P_{i,\,t}$ is then the stock price of company i at time t, while $P_{i,\,t-1}$ is the price of the asset at time t-1.

Assume here that the prices are log-normally distributed so that equation (3.1) will be normally distributed (McDonald, 2009). This gives the current return for assets from time t to t − 1. This means that the sum of daily returns over a period will be equal to the return for this entire period. This can be used to annualize the return. A common assumption when annualizing daily values is that there are 252 trading days in one year (Chan & Wong, 2013). The average annual return is calculated as the arithmetic mean of the periodic logarithmic returns multiplied by trading periods. In our case there are m = 252 periods, or days, per business year.

$$\bar{r}_{i,d} = \frac{1}{n}\sum_{i=1}^{m} r_{i,d} \tag{2.3}$$

$$\bar{r}_{i,ann} = (1 + \bar{r}_{i,d})^{m} - 1 \tag{2.4}$$

Equation (3.3) explains the average return per period for asset i and equation (3.4) the average annual return for asset i (Chan & Wong, 2013). We will use this annual return method for the optimalization by Markowitz theory later.

2.4.2 Risk Premium and Excess Return

For our regression analysis at the end of our paper we use excess return for all factors including our portfolios. Risk premium is the expected additional return an investor gets for allocating capital to a riskier portfolio rather than investing in a risk-free asset. Excess return is the actual difference between return and a riskless investment option.

$$\bar{r}_{RP} = E(r_p) - r_f \tag{2.5}$$

$$r_{excess} = r_p - r_f \tag{2.6}$$

(3.5) defines the risk premium, where E(r_p) is the expected return of a given asset and r_f illustrates the risk-free investment alternative. (3.6) defines the actual excess return, where r_p is the return of a given asset (Bodie et al., 2018).

2.4.3 Capital Asset Pricing Model

The capital asset pricing model (CAPM) is used to calculate expected return for a given investment asset. This method was developed by Sharpe (1964), Lintner (1965) and Mossin (1966). CAPM estimates the expected return for a company based on the company's covariations with the market portfolio, the risk premium in the market and the return on risk-free investment alternatives. The market portfolio is a value-weighted portfolio of all the companies in the market. In a value-weighted portfolio, the portfolio weights of a company are equal to the market value of the company divided by the market value of the entire market. The expected return is estimated with CAPM as in equation 3.7.

$$E_{r_i} = r_f + \beta_i\left[E(r_m) - r_f\right] + \epsilon_i \qquad 2.7$$

Where

$$\beta_i = \frac{\sigma_{r_i,r_m}}{\sigma_m^2} \qquad 2.8$$

Here, E(r_i) is the expected return for companies in, [E(r_m) - rf] is the risk premium in the market and β_i is a measure of covariation between asset i and the market portfolio (Bodie et al., 2018). When we do our regression, we calculate the excess return for all factors and from the analysis we will get a beta for each factor. Before we introduce theory on regression analysis, we will define the other main factors for this paper’s investment risk.

2.4.4 Variance and Standard Deviation

Volatility and standard deviation are of paramount importance for investors and is used for determining risk for a given investment. This is done by analyzing variation in price and return and acts as a measure of risk by investing in a particular asset.

The variance and standard deviation in a sample are calculated based on standard formulas.

$$\sigma_{i,d}^2 = \frac{\sum_{i=1}^{n}(r_{i,d} - \bar{r}_{i,d})^2}{n-1} \qquad 2.9$$

$$\sigma_{i,d} = \sqrt{\frac{\sum_{i=1}^{n}(r_{i,d} - \bar{r}_{i,d})^2}{n-1}} \qquad 2.10$$

$\sigma_{i,d}^2$ is the daily variance of return and $\sigma_{i,d}$ is the daily standard deviation of the return of companies in (Bodie et al., 2018).

To annualize the variance and the standard deviation, the daily variance is multiplied by 252, while the daily standard deviation is multiplied by the square root of 252

$$\sigma_{i,a}^2 = \sigma_{i,d}^2 * 252 \qquad 2.11$$

$$\sigma_{i,a} = \sigma_{i,d} * \sqrt{252} \qquad 2.12$$

(3.11) defines annualized variance and (3.12) the annualized standard deviation (Chan & Wong, 2013).

2.4.5 Covariance & Correlation

The relationship and dependency between two variables can be measured by covariation and correlation. The covariance is a standardized value that indicates the direction of the linear relationship between the variables, and correlation is a non-standardized value that measures both the strength and direction of the linear relationship between the variables. Correlation is therefore better to use when comparing variance between different assets.

$$Cov(r_j, r_k) = \sigma_{j,k} = \frac{\sum_{i=1}^{n}(r_{j,d} - \bar{r}_{j,d})(r_{k,d} - \bar{r}_{k,d})}{n-1} \qquad 2.13$$

$$Corr(r_j, r_k) = \rho_{j,k} = \frac{\sigma_{j,k}}{\sigma_j * \sigma_k} \qquad 2.14$$

where (3.13) is the covariance between the return of asset j and the return of asset k. From the covariance you can calculate the correlation, equation 3.14, between companies that gives a value between -1 and 1. The correlation in return to the assets is calculated by dividing the covariation in return to the companies on the product of the standard deviations of the return to the companies (Bodie et al., 2018). In regression analysis correlation is the method for explaining variance between the dependent variable and the independent factors, but in this calculation, we use covariance matrices to combine more factors in the same model.

2.4.6 Matrix

Matrices and vectors are used to handle information in an efficient way and can easily be used with computer tools. A variance-covariance matrix, $V_{n\times n}$, consists of the variance in the return of the assets along the diagonal in the matrix, and the covariance in the return of the assets symmetrically around the diagonal. They are symmetrical around the diagonal since $\text{Cov}(r_k, r_j) = \text{Cov}(r_j, r_k)$. We construct a variance-covariance matrix according to the following principle.

$$V_{3*3} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_1^2 & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_1^2 \end{pmatrix} \quad 2.15$$

Based on the variance-covariance matrix, we can construct a correlation matrix. We must then go through a matrix that consists of all the products of the standard deviations of the assets returns.

$$S_{n*n} = s_n^T * s_n \quad 2.16$$

where T is the notation for the transposed of a vector or matrix.

Here we have an example of the Variance-Covariance matrix calculation for a dataset with 3 variables.

$$S_{3*3} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix} * \begin{bmatrix} \sigma_1 & \sigma_2 & \sigma_3 \end{bmatrix} = \begin{pmatrix} \sigma_1 * \sigma_1 & \sigma_1 * \sigma_2 & \sigma_1 * \sigma_3 \\ \sigma_2 * \sigma_1 & \sigma_2 * \sigma_2 & \sigma_2 * \sigma_3 \\ \sigma_3 * \sigma_1 & \sigma_3 * \sigma_2 & \sigma_3 * \sigma_3 \end{pmatrix} \quad 2.17$$

By dividing the elements of the variance-covariance matrix, $V_{n\times n}$, element by element on the elements of the matrix $S_{n\times n}$, we get the correlation matrix, K_{n*n}.

$$K_{3*3} = \begin{pmatrix} 1 & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & 1 & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & 1 \end{pmatrix} \quad 2.18$$

The elements in the diagonal will be equal to 1, because this shows the assets correlation to itself. Assets with low correlation coefficients indicate that they are suitable to include in a portfolio in order to obtain a diversification gain (Bodie et al., 2018; Sydsæter & Hammond, 2008). This is especially helpful when constructing the minimum variance portfolio, where we want to take advantage of correlation to archive a lower risk than the least risky asset in our portfolio.

In the example variance-covariance equation 3.17 and correlation matrix equation 3.18, we have a dataset with 3 variables. For our dataset, 1 period of covariance matrix look like table 2 where we have 5 variables to calculate one periods annualized variance-covariance matrix.

Table 2: Annualized Variance-Covariance Matrix

	DIA	QQQ	SPY	SUSA	VUG
DIA	0,0136	0,0114	0,0128	0,013	0,0115
QQQ	0,0114	0,0296	0,0181	0,0192	0,0272
SPY	0,0128	0,0181	0,015	0,0156	0,0174
SUSA	0,013	0,0192	0,0156	0,0166	0,0185
VUG	0,0115	0,0272	0,0174	0,0185	0,0257

2.4.7 Portfolio Expected Return

A portfolio is composed of several assets with different expected returns, $E(r_i)$. Expected return for the entire portfolio, $E(r_p)$, is the weighted average of expected return for each individual company, where the weights, w_i, are the proportion of total investment invested in companies i (Bodie et al., 2018).

$$E(r_p) = \sum_{i=1}^{n} w_i * E(r_i) \quad 2.19$$

where $E(r_i)$ is the expected return for the company estimated in equation (3.7).

We use this method for both the minimum variance portfolio and the maximum Sharpe-Ratio portfolio where we combine 50.000 portfolios with different weights and calculate their expected returns to create an efficient frontier.

2.4.8 Portfolio Variance & Standard Deviation

The variance in return of a portfolio is a weighted sum of variance and covariance in the return on the companies included in the portfolio (Bodie et al., 2018).

$$Var_d\left(r_p\right) = \sigma_p^2 = \sum_{i=1}^{n} w_j * w_k * \sigma_{j,k} \tag{2.20}$$

When the covariance coefficients are calculated based on daily observations, σ_p^2 will be the portfolio's daily variance. We find the portfolio's annual variance by multiplying the daily variance in equation (3.20) by 252. The annual standard deviation for the portfolio is the square root of the annual variance.

$$\sigma_{a,p}^2 = \sigma_p^2 * 252 \tag{2.21}$$

$$\sigma_{a,p} = \sqrt{\sigma_{a,p}^2} \tag{2.22}$$

where equation (3.21) defines the annual variance and equation (3.22) defines the annual standard deviation (Chan & Wong, 2013).

2.4.9 Diversification

The risk one asset or a portfolio is exposed to can be divided between systematic and idiosyncratic risk. Idiosyncratic risk is the risk that accompanies each individual company. This can, for example, be a risk of poor management, danger of strikes, opportunities for outdated products and so on. By distributing your investment in several companies, you will be able to reduce this company-specific risk. Systematic risk is the risk that is affected by common risk factors, and these cannot be diversified away by investing in several companies. Systematic risk is therefore also called undiversifiable risk (Bodie et al., 2018). This is illustrated in figure 2, where the total risk (blue line) is a combination of systematic and unsystematic risk.

Diversification will not be an issue for our study as we use ETFs, which is a composition of multiple assets and therefore has reduced unsystematic risk.

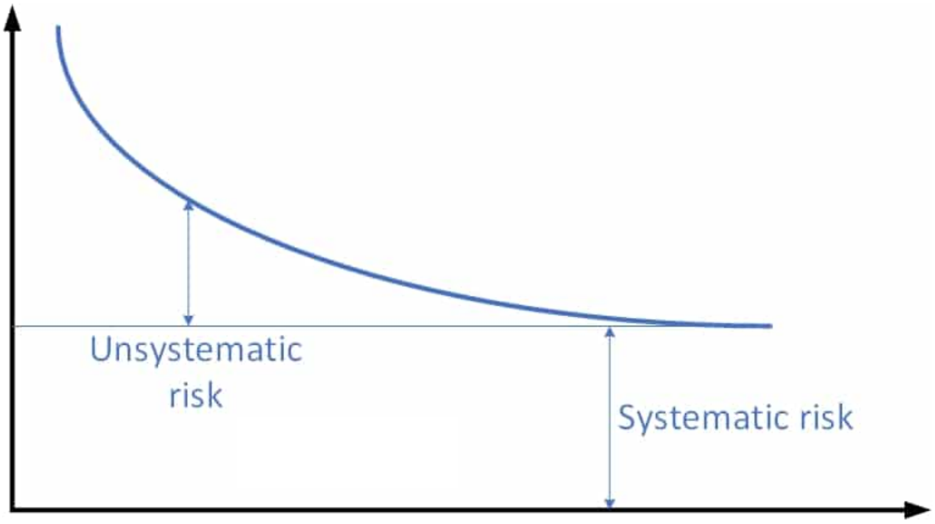

Figure 2: Diversification Illustrated.

2.5 Regression Analysis

In this paper we will end with a regression analysis to evaluate performance and try explaining some of our excess return. In this chapter we will discuss the purpose of a linear regression and why we will use it in our study. We start by introducing the theoretical concept before we implement the theory on our data and explain how we methodically approach it to our analysis.

2.5.1 Linear Regression

The purpose of a linear regression model is to explain and evaluate the covariance between two or more variables. As long as there is a covariance between the variables, the dependent variable (Y) can be explained as a function of the independent variables (X) (Albrigtsen, 2007).

$$Y = \hat{\alpha} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \ldots \widehat{\beta_k}X_k + \epsilon \qquad 2.23$$

Where,

Y: Observed values on the dependent variable.

$\hat{\alpha}$: predicted constant (crossing line between the y-axis and the regression line). Y’s value when $X_i = 0$.

$\widehat{\beta_{1\rightarrow k}}$: Predicted regression coefficients. Expresses the change in Y given 1 increase in X, ceteris paribus.
ϵ: Error term or residual error. The amount of Y that is not explained by the dependent variables.

The regression line as a result is called the OLS-regression line because the equation is solved using, Ordinary Least Squares. OLS minimize the vertical squared lines between the observed and predicted values, so that the line is the best fit for all the plotted data. When this is done, you also minimize the prediction error term, ϵ (Amundsen & Statistisk, 1980). In figure 3, we have illustrated the method where the green line is the regression equation and the black spots are the actual data, in our example monthly return of the chosen portfolio, the blue lines are the result from the OLS, squared variances.

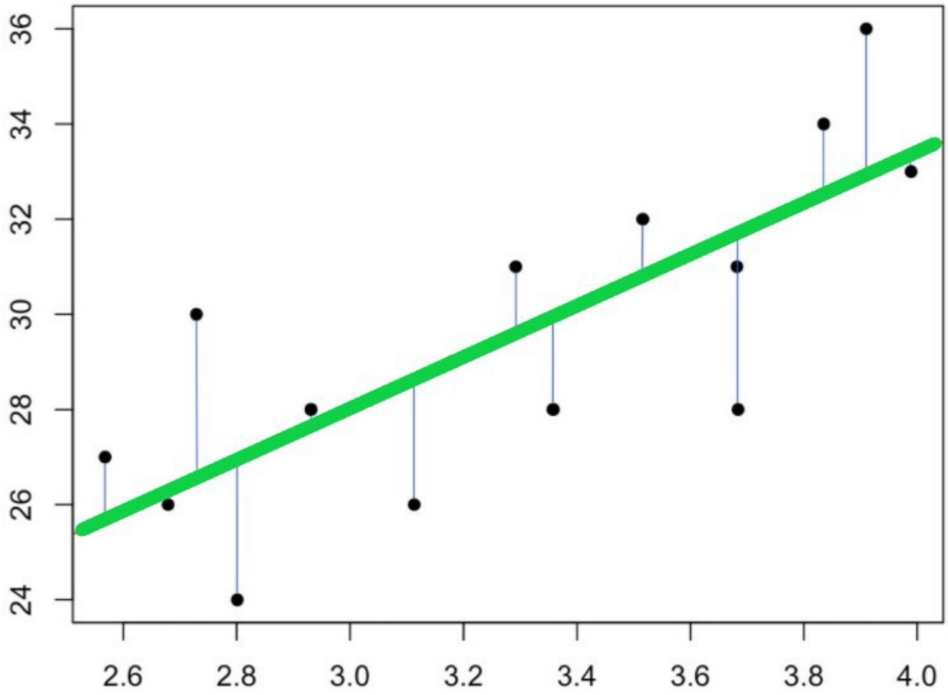

Figure 3: The OLS-Regression (Thieme, 2021)

In this study we will focus on the R^2 and p-value form the regression.

2.5.2 R-Squared

The R-squared (R^2) is used in the regression model to explain how much of the variance in the dependent variable is explained by variance in the independent variables. We will define our dependent variables and our independent variables later.

If we imagine that all our observations are two-parted. One explained part and one unexplained, $y_i = \hat{y}_i + e_i$, we can then define:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad 2.24$$

Equation 3.24 is the total sum of the squared variances, TSS (Total Sum of Squares)

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad 2.25$$

Equation 3.25 is the explained part of the variances, ESS (Explained Sum of Squares)
While $\sum_{i=1}^{n} e_i^2$ is the part of the variance to the dependent variable, that cannot be explained by the model, RSS (Residual Sum of Squares).
R^2 for the regression is the "Total Sum of Squares" that is explained by the model and can be defined as:

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \qquad 2.26$$

The value of R^2 must therefore by definition, be between 0 and 1. The higher R^2 is, the better we can say that the variance in the model explain the variance in the data.

2.5.3 p-Value

The p-value is used to validate the statistical significance to the factors in the regression model. The p-value gives us the lowest critical significance level of alpha (α) we can have and still reject the null-hypothesis. The null hypothesis is that the factor coefficient in the regression is not statistically different from zero, $H_0 = \beta = 0$ against the alternative hypothesis, $H_1 = \beta \neq 0$ where the factor coefficient is not equal to zero.
Example: If $y = \beta_0 + \beta_1 x + e$ and a significance level of $\alpha = 0,05$ (5%) a p-value of $\beta_1 = 0,028$ will result in a rejection of the null hypothesis $\beta = 0$ and we will accept the alternative hypothesis $\beta \neq 0$, we therefore have a statistical significant covariance between the dependent variable (Y) and the independent variable (X), and only a 2,8% chance of making

a type-I error, to reject the null hypothesis when it is actually true (Amundsen & Statistisk, 1980).

2.5.4 The Fama French Carhart Model

In 1996, Eugene Fama and Kenneth French expanded on the CAPM theory by adding size risk and value risk factors to the return of the market factor in CAPM. They figured from research that value stocks tend to outperform growth stocks and small-cap stocks tends to outperform large-cap stocks (Fama & French, 1996).

If we consider the CAPM model from chapter 2.4.3, equation 2.7:

$$E(r_i) = r_f + \beta_i[E(r_m) - r_f] \qquad 2.27$$

If we then denote $E(r_m) - r_f$ as Mkt (Excess Market return) and add the two factors from Fama French. We get:

$$E(r_i) = r_f + \beta_1 * Mkt + \beta_2 * HML + \beta_3 * SMB + \epsilon \qquad 2.28$$

This model considers the fact that value and small-cap stocks outperforms the market on a regular basis. By adding these two additional factors, the model will be adjusted for this outperforming tendency, which is thought to make it a better tool for evaluating portfolio performance (Hayes, 2021).

The Carhart 4 factor model (also known as Fama French Carhart factor model) was introduced by Mark Carhart in (1997) a few years after Fama and French introduced the first 3 factors. Carhart used the cross-sectional momentum factor to show that it was distinct from Fama and French 3 factors and that it improved the explanatory of multifactor models, aimed at explaining mutual funds' performance, significantly.

If we add the momentum factor to the model, we get:

$$E(r_i) = r_f + \beta_1 * Mkt + \beta_2 * HML + \beta_3 * SMB + \beta_4 * MOM + \epsilon \qquad 2.29$$

There are different ways to calculate the factors, all depending on what type of assets you want to price, or in our case, what our excess returning portfolios consists of. Since we look at different markets within the American stock market, we can collect our factors directly from one of the founders of the model, Kenneth R. French's website (French, 2022). On this website he has created factors to different markets and the ones we will be using in this study is:

Mkt: The excess market return factor is created by constructing a market portfolio and subtracting the risk-free rate from the monthly return of the portfolio. The market portfolio consists of all NYSE, AMEX and NASDAQ firms.

HML: The High Minus Low factor is the average return of two value portfolios minus the average return on two growth portfolios. Small Value portfolio consists of the bottom 30% of book/equity-ratio stocks listed on the NYSE, AMEX and NASDAQ. The Big Value portfolio consists of the top 30%. For the Growth portfolios it is the top and bottom 30% of the stocks that have the largest and smallest investment, measured in change in total assets, the fiscal year t-2 to t-1.

$$HML = \frac{\frac{1}{2}(Small\ Value + Big\ Value)}{\frac{1}{2}(Small\ Growth - Big\ Growth)} \qquad 2.30$$

SMB: Small minus Big is the average return of nine small stock portfolios, lowest 30% measured in market equity, minus the average return on nine big stock portfolios, measured at the top 30%.

$$\begin{aligned} SMB_{\left(\frac{B}{M}\right)} &= \frac{1}{3}(Small\ Value + Small\ Neutral + Small\ Growth) \\ &\quad - \frac{1}{3}(Big\ Value + Big\ Neutral + Big\ Growth) \\ SMB_{(OP)} &= \frac{1}{3}(Small\ Robust + Small\ neutral + Small\ weak) \\ &\quad - \frac{1}{3}(Big\ Robust + Big\ Neutral + Big\ Weak) \end{aligned}$$

$$SMB_{(INV)} = \frac{1}{3}(Small\ Conservative + Small\ Neutral + Small\ Aggressive) - \frac{1}{3}(Big\ Conservative + Big\ Neutral + big\ Aggressive)$$

$$\boldsymbol{SMB} = \frac{1}{3}(SMB_{\frac{B}{M}} + SMB_{OP} + SMB_{INV}) \qquad 2.31$$

MOM: The momentum factor is the average return on the two high prior return portfolios minus the average return for the two low prior return portfolios. The portfolios are constructed by combining five portfolios, two portfolios based on market equity and tree based on prior return. The stocks breakpoint for high/low on size is the median NYSE market equity. The prior (2-12) return breakpoints are the 30th and 70th NYSE percentiles.

$$MOM = \frac{1}{2}(Small\ High + Big\ High) - \frac{1}{2}(Small\ Low + Big\ Low) \qquad 2.32$$

In our study we will do one regression analysis to each of our 5 portfolios with the portfolio being the dependent variable and the Fama French Carhart 4 factors being the independent variables. This analysis will be done to see how much of the portfolios return can be explained by the 4 factors. We will use the p-value to determine whether we have a significant intercept (α), meaning weather or not the portfolios excess return is statistical significantly different from zero.

Keep in mind that even though we get a significant excess return, there are some weaknesses to regression analysis. This can be discussed when we look at the assumptions for multiple linear regression models (CorporateFinanceInstitute, 2015).

1. The dependent and independent variables show a linear relationship between the slope and intercept.
2. The independent variables are not random
3. The value of residual error is zero.
4. The value of residual error is constant across all observations.
5. The residual error values follow the normal distribution.

6. Independent variables should show a minimum correlation with each other.

The last assumptions are one of the most important and most difficult to eliminate when we want to explain portfolios return. If the independent variables are highly correlated with each other, it is difficult to assess the true relationship between the dependent and independent variables (CorporateFinanceInstitute, 2015).

3 Data and Methodology

This chapter is divided into two parts. One for the data used in this study and then the 5 portfolio allocation strategies we have chosen to compare. The methods used in this study will mostly follow Markowitz Modern Portfolio Theory (1952) and Tan et al. (2019) for the Machine Learning part. For the calculations and simulations, we have used a combination of excel to clean datasets and R to code our portfolios and to do the actual calculations and simulations.

The method for the first 4 portfolios was divided into 3 parts:

1. The data is collected, combined and cleaned.
2. The dataset is split into training periods. Training periods are used to calculate the average return and variance which are then used to create the efficient frontier.
3. The training periods are being used to construct portfolios.

The method for the Random Forest portfolio was divided into 5 parts:

1. The data is collected, combined and cleaned.
2. The dataset is split into periods, one for training of the model and one for trading. The training data is used to train the model and the trading data is used to test the model on unknown data.
3. The function and response variable for the model is constructed. Function consists of entry values that are needed in the construction of the model to be used in prediction.
4. The model is trained with the training data.
5. The trained model is used on unknown dataset, trading data, and the predictions are then used to construct portfolios.

Continuing this paper, we will illustrate the strategies and portfolio weights using the abbreviations given in table 3.

Table 3: List of Strategies and Abbreviations

Number	Strategy	Abbrevation
1.	Buy and Hold Portfolio	B&H
2.	Naive 1/N with rebalancing (benchmark)	1/N
3.	Minimum Variance Portfolio	MV
4.	Tangency Portfolio (Max Sharpe-Ratio)	SR
5.	Machine learning (Random Forest)	RF

3.1 Data

This study focuses on 5 distinct ETFs, all listed on the American stock market and which tracks different indexes in the American stock market.

1. DIA – The Dow Jones Industrial Average is composed of 30 large-cap ("blue-chip") stocks and seeks to provide investment results that, before expenses, correspond generally to the price and yield performance of the Dow Jones Industrial Average (Index). It is being recognized as the only index composed of companies that have sustained earnings performance over a significant period of time (SSGA, 2022).
2. QQQ – Invesco QQQ ETF tracks the Nasdaq-100 Index, meaning it tracks the 100 largest non-financial companies listed on the Nasdaq, based on market cap. As of March 2022, it is rated as the best performing large-cap growth fund (1 of 317), based on total return over the past 15 years and is the 2nd most traded fund (Invesco, 2022).
3. SPY – The S&P 500 is one of the main benchmarks of the U.S. equity market and indicated the financial health and stability of the economy. The SPDR S&P 500 ETF Trust (SPY) tracks the performance of this index. SPY is recognized as the first ETF in the U.S market and by many investors the best investment for non-experts, as it has yielded an average of over 10% return since inception (Nickolas, 2022).

4. SUSA – The iShares MSCI USA ESG Selected ETF seeks to track the performance of an index composed of U.S. companies that have positive environmental, social and governance characteristics as identified by the provider. As part of the screening they exclude and limit exposure to certain controversial companies or sectors such as firearms, alcohol and gambling (BlackRock, 2022).
5. VUG – The Vanguard Growth ETF seeks to track the performance of the CRSP U.S. Large Cap Growth Index. It limits its investment by not exceeding 10% of total outstanding voting securities of any issuer or purchase securities of any issuer is, as a result, more than 5% of the fund's total asset found be invested in that issuer's security, unless it is necessary to track the index (The Vanguard Group, 2022).

We have chosen to use only ETFs with American stocks. This is because of the markets liquidity and to test the proclaimed strong market efficiency. It will also be easier to compare our study to others, since the American stock market is one of the most analyzed markets in financial economics.

We have used Microsoft Excel to sort some of our data but, mostly used the programming tool R. We used the package "tidyquant 1.0.3" by Matt Dancho (2021) to collect daily return data for each ETF. This package collects data from Yahoo Finance. For the Random Forest model, we have used the package "tidyquant" to collect data, numerous packages to calculate the variables and lastly, we use the package "randomForest" to execute the Random Forest algorithm.

For the 3 last strategies we use rolling window to calculate next periods portfolio allocation. For the machine learning strategy we use out-of-sample rolling window with a training period of 5 years (2011-2016). The rolling window consist of daily returns for 1-year period at a time, about 252 observations (black in figure 4). This period is used to calculate the yearly average return and covariance matrix for MV- and SR-portfolio. We then compute 50 000 random weighted portfolios to create the efficient frontier witch we then extract the optimal weights for the Minimum Variance and max Sharpe-Ratio portfolio for the next period of 1 month (yellow in figure 2). We continue this process for 6 years (2016-2021) until we have 60 portfolios with different weights. We then multiply the actual returns for the ETFs by their weights to get the portfolio risk and return. We use the last 6 years for our analysis because they are more relevant for this research and our comparison.

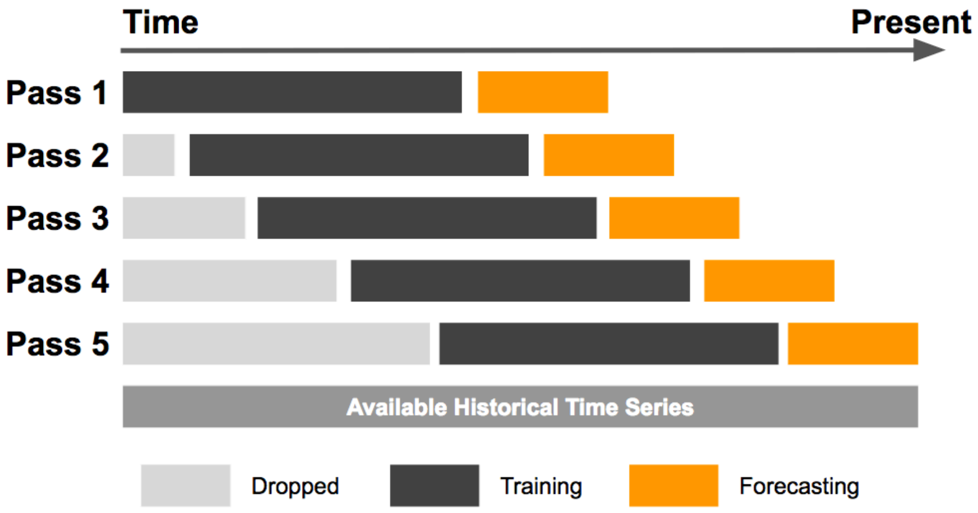

Figure 4: Sliding Window (Yang, 2018)

The RF portfolio has a training period consisting of five years of monthly data, and a forecasting period (test period) of one month. This last month gives us the probability of the asset rising in price over the next period, in this case the next month. This is done for all periods during the six year back testing timeline, by looping the Random Forest algorithm.

3.2 Buy and Hold

The efficient market theory suggest that it would be impossible for an investor to outperform the market using investment tools such as fundamental or technical analysis (Cohen & Cabiri, 2015). In this case, he or she should therefore put all his assets in the buy and hold portfolio as it is easier to adopt and cheaper in terms of transaction cost compared to other investment strategies (Shilling, 1992). This strategy involves no optimalisation or estimation and completely ignores the data. In fact, this portfolio will only have transactions at the start of the period t=0 and at the end of the last period, when the portfolio is liquidated. For this reason, it would be the best portfolio, compared to the naive 1/N portfolio, if we include transaction costs and an efficient market. Another reason to opt for this strategy is the tax benefits. Tax on income is only calculated when you either have a positive yield on investments or deducted when you have a net loss. In other words, there are no income tax if you do not sell any assets, given you do not have a share savings account. These accounts allow you to buy and sell certain assets without triggering tax.

Warren Buffet, by many stated to be the greatest investor of all time, has at numerous occasions said that his favorite holding period is “forever”. This is referring to the buy and hold strategy. There are however three problems with Warren Buffets comment about the Buy and Hold strategy and the strategy itself (Doroghazi, 2020).

1. During a bear market period, it can take an agonizingly long time for the market to recoup the losses. The Dow Jones Industrial Average Index (DIJA) did not return to the 1929 high until 1954. In other words, it took the index a quarter of a century to get back to an all-time high. Another example is the DJIA all-time high in 1966 of 1000 points. The market then went bear and did not hit 1000 again until late 1982, but when factoring in the high inflation of the 70s, the DJIA did not return to the 1000 points mark until 1991. The dot-com bubble of the 2000 saw the DJIA lose its all-time high of 11.700, (as low as 6547 in 2008, 45% off the 2000 high), before closing above in 2012. It has since not been below the high-mark of 2000, but there is certainly no guarantee it will not happen in the future. Recently, the COVID-19 pandemic saw the DJIA lose almost 3000 points (-12.9%). This crash however did not last longer than 5-6 month before the DJIA set a new all-time high. Many investors point to the Federal Reserve for this as they answered the crash with a 1,5% cut in federal funds rate, quantitative easing (QE) meaning they purchased enormous amounts of security debt, and lending money to securities firms, as well as more (Wessel, 2021). To sum up, all these three bear market periods (except COVID-19) have resulted in half a century of agonizing pain and volatility, just to get even (Doroghazi, 2020).
2. Although Buffet himself have claimed that his favorite holding period is forever and that it is impossible to time the market according to him. In 1987 Buffet liquidated the retirement portfolios of Berkshire Hathaway’s employees, right before a market crash in October and if you read the quarterly reports of Berkshire Hathaway, you can see that Berkshire Hathaway is buying and selling frequently throughout the year. This has led to some allegations that Mr. Buffet and his company are in fact trading after the Quality factor, intentionally or not (Patel, 2018).
3. A key drawback with the buy-and-hold is that over time, there is a chance that one asset grows a lot more in value than the other assets, and after a while the entire portfolio’s performance is mainly driven by one asset.

The portfolio weights evolve according to the relationship:

$$w_{it} = w_{i(t-1)} \frac{r_{it}}{r_{p_t}} \tag{3.1}$$

Where t denotes the time period, $r_{it} = \frac{x_{it}}{x_{t-1}}$ is the one period price relative for asset-i where x_{it} is the stock price, and $r_{pt} = \sum_{j=1}^{m} w_{j(t-1)} r_{jt}$ is the portfolio return.

As the period progress the returns of each individual ETF can be seen in figure 4 and the weights shift to be more weighted on the assets with higher return while the lower returning ETF´s lose percentage of the total weight of the portfolio.

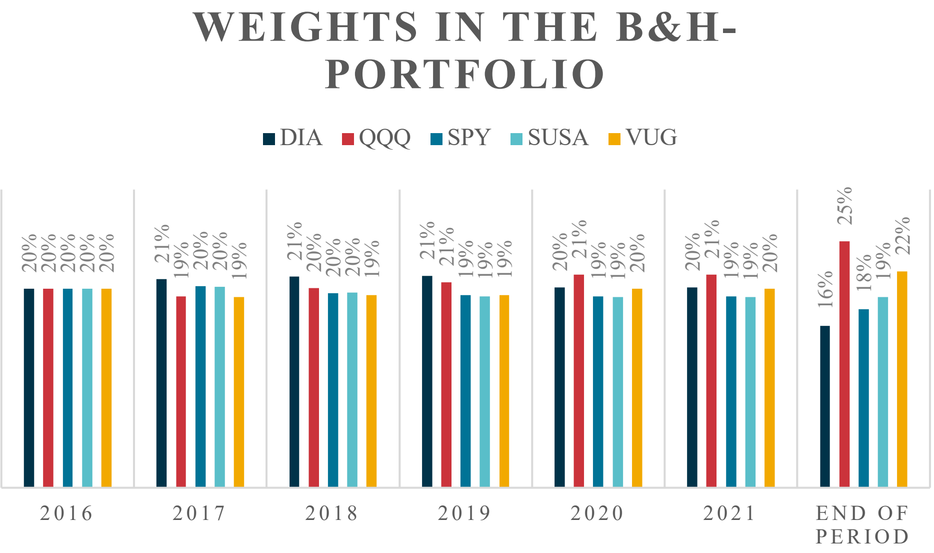

Figure 5: Weights in the B&H Portfolio

3.3 Naive 1/N

The naive 1/N strategy is an equally simple portfolio as the buy and hold portfolio, the only difference being that at the beginning of each month we rebalance the portfolios assets to be equally weighted again. It also, does not involve any optimalization or estimations and completely ignores the data. For the start of the period t=0 we have the same wights as the buy and hold portfolio, 1/5. At the first trading day of each month, we rebalance the portfolio to a weight vector which divides the accumulated wealth equally among the ETFs. This process will continue until the end of period t=72.

So, for each rebalance we get, $w_t^{ew} = 1/N$ in each ETF at time t, shown in figure 6.

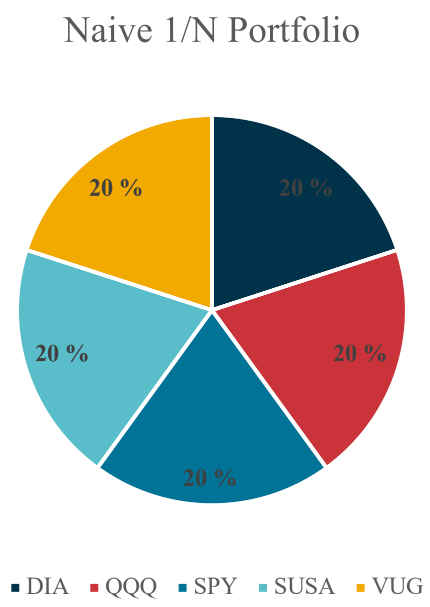

Figure 6: Naive 1/N Portfolio Wights

Unlike the buy and hold strategy this portfolio will not have the same advantages concerning taxes, as the portfolio will be rebalanced monthly witch implies realization of returns, both gains and losses. This amount of transaction will also add up to the total cost of the strategy and might in practice eat up some of the return (even though we have already mentioned its marginal effects). The biggest advantage compared to the buy and hold strategy is the monthly rebalancing, making sure one of the portfolios assets do not become too large compared to the other assets. This could as mention, lead to a portfolio mainly driven by one assets volatility.

If we look at the general formula for the variance of a portfolio with size n assets:

$$\sigma_p^2 = \sum i = 1n\sum j = 1n\ w_i w_j Cov\left(r_i, R_j\right) \tag{3.2}$$

Where w_i and w_j are the investment weights in asset I and j respectively. r_i and r_j are the returns on asset i and j, respectively, $Cov(r_i,r_j)$ represents the covariance between returns on asset i and j. In a naive equally weighted portfolio we have $w_i=w_j=1/N$, so the equation can be rewritten as:

$$\sigma_{p,1/n}^2 = 1n\sum i = 1n1n\sigma_i^2 + \sum i = 1_{i\neq j}n\sum j = 1n1n^2Cov\left(r_i, r_j\right) \tag{3.3}$$

In equation 4.3 there are n variance terms and n(n-1) covariance terms. If we then let the average variance and average covariance be

$$\sigma^2_{n,1/n} = 1n\sum i = 1n\sigma_i^2, \tag{3.4}$$

$$Cov_{n,1/n} = 1\mathrm{n}(n-1)\sum i = 1_{i\neq j}n\sum j = 1n\ Cov\left(r_i, r_j\right) \tag{3.5}$$

Then equation 4.5 can be expressed as

$$\sigma^2_{p,1/n} = 1n\sigma_n^2 + n - 1nCov_{n,1/n} \tag{3.6}$$

From equation 4.6 we can clearly see that when portfolio size (n) increase, the first term on the righthand side tends to zero while second term tends to the average covariance (as (n-1) /n tends to 1). This equation also describes the portfolio variance for an equally weighted portfolio (Tang, 2004) and thus we can write the general formula for variance in an equally weighted portfolio as:

$$E\left(\sigma_n^2\right) = 1n\sigma_N^2 + n - 1nCov_N \tag{3.7}$$

Where n is the number of assets in the portfolio, n=1,2,….,N, N is the number of assets in the population and $E\left(\sigma_n^2\right)$ is the expected portfolio variance with portfolio size of n, σ_n^2 is the expected variance of all assets in the population and Cov$_\mathrm{N}$ is the expected covariance of all asset in the population.

From this we can see that when the portfolio size n increases, the first term on the right-hand side in equation 4.7 will become smaller and the term on the left-hand side will become larger and tends to the Cov$_\mathrm{N}$. This means that the first term of risk is diversifiable (firm-specific risk) while the second term is the non-diversifiable risk (systematic risk).

3.4 Minimum Variance

Modern portfolio theory (MPT) is a selection method for assembling different assets with the intent to maximize their overall return and minimizing risk for an investor. The theory was pioneered by, Markowitz (1952). A crucial component in MPT is diversification since investments either are associated with high risk and high return or low risk and low return. In this paper Markowitz argue that a combination of multiple assets with both high and low risk can result in a portfolio that yields better results given the individual risk tolerance (Markowitz, 1952). This work within modern portfolio theory was later awarded a Nobel prize.

It was however, not until fast computers were advent in the 1970s, that this modern approach to portfolio management were implemented in practice. Prior to this, investors were thought to have made inefficient, unsophisticated portfolio selections (Sotiropoulos & Rutterford, 2019). An assumption under the CAPM and therefore MPT, is the assumption that all investors are risk averse, meaning they will always choose the optimal portfolio with the highest return if the portfolios have the same level of risk. The most common measure for risk in a portfolio is the variance of returns. This application of mean-variance model to portfolio selection problem laid the foundation for MPT. The main insight of this theory is simple: "if individual security risk is captured by the expected variance of returns, portfolio risk requires a set of variances and covariances in order to be fully described. In other words, when it comes to the analysis of portfolio risk, one needs to take into account not only the individual components' risks but also their interactions. The optimization solution leads to a hyperbola depicting all the possible maximum returns for a given level of risk in the risk-expected return space. From all possible combinations in the hyperbola curve, the optimal portfolio is the one that maximizes the so-called Sharpe ratio, which is equal to the expected excess return of the portfolio relative to the risk-free asset, divided by the expected standard deviation of the portfolio." (Sotiropoulos & Rutterford, 2019). The other portfolio derived form this is the minimum variance portfolio which will have the lowest amount of risk taken. In this section we will lay the theoretical grounds for both portfolios but only explain the method for how we computed the minimum variance portfolio. In the following chapter we will describe how we created the optimal portfolio. The theory of the mean-variance relationship between risk and return will however remain the same for both portfolio selections.

There are at least two challenges with Markowitz's theory.

1. It is not that straight forward in practice to create a mean-variance optimal portfolio selection. The optimal solution for example is very sensitive to initial assumptions

with regard to investors expectations and future returns. This can often lead to significant estimation errors since optimized portfolios are rarely optimal in practice.

2. More importantly, investors were not helpless, and they did not at all lack investment advice on how to diversify their portfolios before the rise of MPT. The practice of spreading risk across a number of different assets worldwide was widely and consistently promoted, at least within the UK financial community, at least from the 1970s (Sotiropoulos & Rutterford, 2019). Investors where already informed about the advantages of spreading their wealth across different assets, as financial advisors and analysts offered detailed recommendations on how to best combine investments to enhance yield without increasing portfolio volatility. Several books, pamphlets, magazines, and newsletters though investors on the theory of naïve diversification, that is, equally weighting their portfolios across risky assets, like explained in our naïve portfolio.

The variance of the returns of the portfolio can be written from MPT as:

$$Var_d(r_p) = \sigma_p^2 = \sum_{i=1}^{n} w_j * w_k * \sigma_{j,k} \tag{3.8}$$

Were all asset weights are summed and w$_x$ are the weights of the different assets and $\sigma_{j,k}$ denotes the covariance between all assets. Since $\sigma_{j,k} = Cor\ \sigma_j\ \sigma_k$ the theory implies that portfolio variance can be reduced by choosing asset classes with a low or negative correlation. We'll now show how we calculated the minimum variance portfolio with our ETF's with monthly rebalancing.

To estimate the minimum variance portfolio, we start by collecting daily return for each ETF for a period of one year. For the first period t=0 we use data from 01.01.2015 to 31.12.2015 to first calculate the mean daily returns for each ETF, we then calculate the covariance matrix and annualize it by multiplying all numbers by 252 (number of trading days in a year).

The weights (w$_{MV}$,i,…, w$_{MV}$,N) are given as (Finance, 2017):

$$w_{MV} = \frac{\sum^{-1} 1}{\mathbf{1} \sum^{-1} \mathbf{1}} \tag{3.9}$$

Where **1** is a column vector of ones. The expected return was calculated as

$$\mu_{MV} = \mu' w_{MV} = \frac{\mu' \sum^{-1} \mathbf{1}}{\mathbf{1} \sum^{-1} \mathbf{1}} \tag{3.10}$$

Where μ is the vector of expected returns. The portfolio minimum variance equals

$$\sigma_{MV} = w'_{MV} = \frac{\mathbf{1}}{\mathbf{1} \sum^{-1} \mathbf{1}} \tag{3.11}$$

We then create 50 000 portfolios with random weights. All these 50 000 portfolios will have different variance, return and Sharpe-ratio. We then extract the portfolio with the lowest variance, the minimum variance portfolio. This portfolio will have the optimal weights for lowest risk and since we do not include risk free rate, the Sharpe-ratio of the portfolio will be portfolio returns divided by portfolio risks.

Since all investors should seek to maximize return at any given risk, the minimum variance portfolio will be at the minimum variance frontier (figure 7) with the lowest possible variance out of all portfolios.

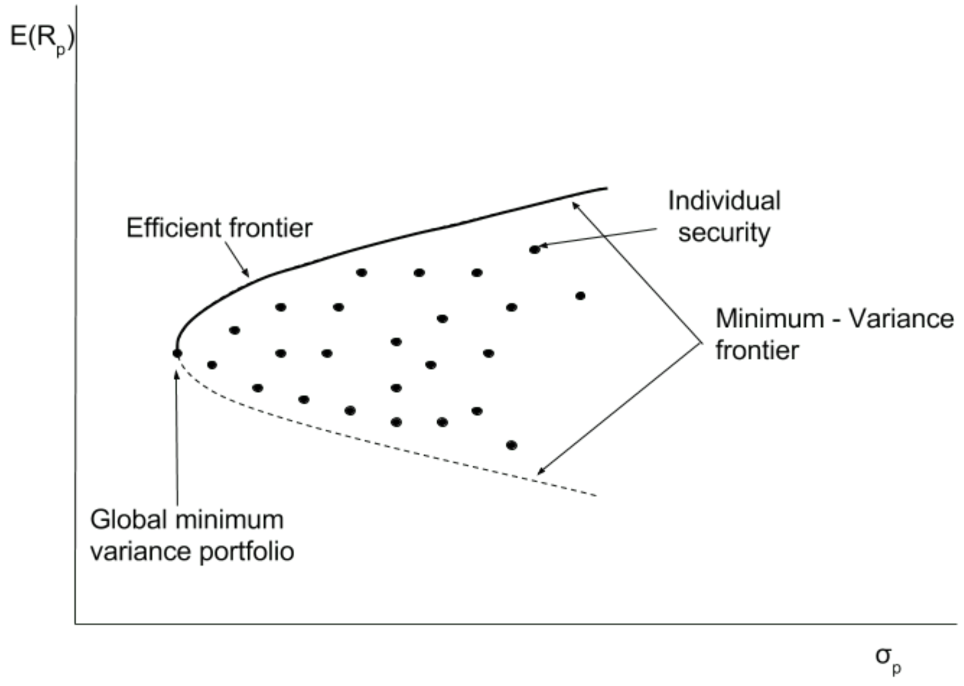
E(R_p)
Efficient frontier
Individual security
Minimum - Variance frontier
Global minimum variance portfolio
σ_p

Figure 7: Efficient Frontier MV

3.5 Max Sharpe-Ratio Portfolio

First introduced by William F. Sharpe who also was a co-founder of the CAPM model (Sharpe, 1964). Sharpe-ratio is a measure of excess return over the risk-free rate per unit of volatility. One of the assumptions in CAPM is that all investors should seek to maximize return and at the same return the investor should always prefer the asset allocation with the lowest volatility. This means that a rational investor always chooses the portfolio with the highest Sharpe-Ratio, because this gives the best expected return at the least amount of risk (Bodie et al., 2009). In portfolio terms we can write the equation for Sharpe-Ratio as,

$$SR = \frac{E(r_i)}{\sigma_i} \qquad 3.12$$

Here, the E(r_i) is the expected excess return of the portfolio, since we however do not include risk free rate in this study, the excess will be the same as E(r_i) and σ_i is the portfolios standard deviation.

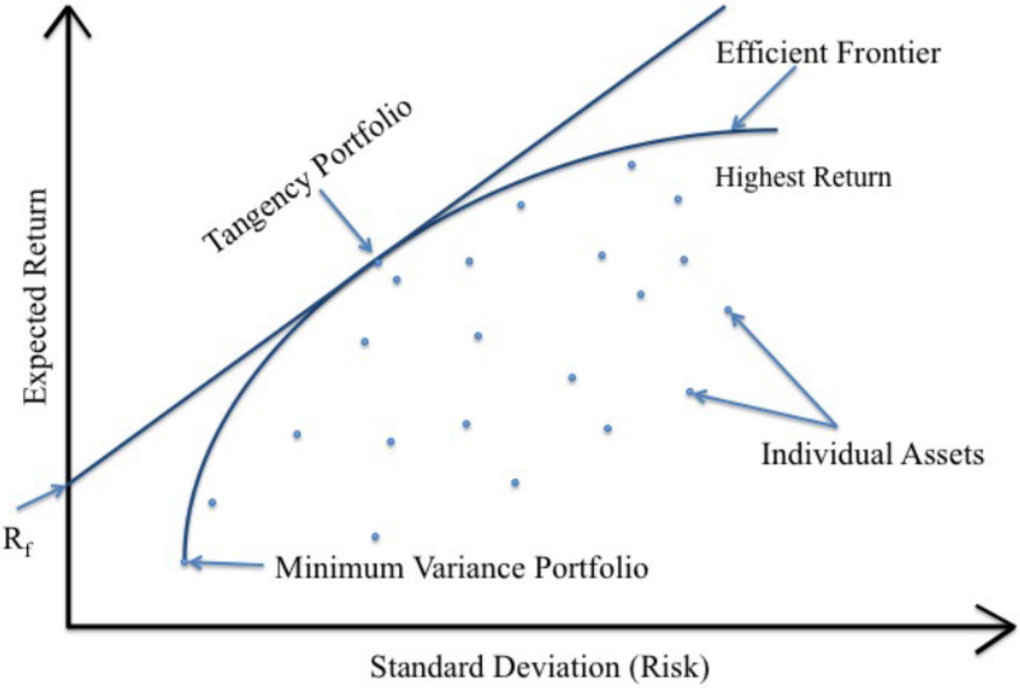

Figure 8: Efficient Frontier Curve SR

To estimate the max Sharpe-Ratio portfolio we start by collecting daily return for each ETF for a period of one year. For the first period t=0 we use data from 01.01.2015 to 31.12.2015

to first calculate the mean daily returns for each ETF, we then calculate the covariance matrix and annualize it by multiplying all numbers by 252 (number of trading days in a year). We then create 50 000 portfolios with random weights. All these 50 000 portfolios will have different variance, return and Sharpe-ratio. We then extract the portfolio with the highest return at any given risk, the max Sharpe-Ratio portfolio. This portfolio will have the optimal weights for per unit of risk and since we do not include risk free rate, the Sharpe-ratio of the portfolio will be portfolio returns divided by portfolio risks.
A risk-averse investor will always pick a portfolio on the efficient frontier (figure 8) since these portfolios give the investor the highest return for each level of risk. The tangency portfolio (optimal portfolio) will have the highest Sharpe ratio of all portfolios, and thus, is the most optimal portfolio.

3.6 Random Forest Portfolio

The content of this chapter is based on the book introduction to statistical learning, second edition, written by Gereth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013). All theoretical concepts and methods for constructing the portfolio based on the Random Forest algorithm is presented in the following chapters below.

3.6.1 Introduction to Machine Learning

Prediction of data from the financial market is challenging due to the non-stationary nature of the data, since the mean values, variance, and covariance changes over time. The data are often chaotic, non-linear and contain noise (Henrique et al., 2019). Technological advances have made it possible to analyze large amounts of historical data using computer software, and at the same time find connections that is not clearly visible for the human eye. This use of intelligent models is often studied under the title machine learning. Machine learning techniques integrates computer software with artificial intelligence and seeks to extract patterns from historical data.

Machine learning is based on statistical learning. From this we have input variables referred to as predictors, and output variables referred to as response value. The input values are denoted by a matrix *X* consisting of p vectors $X_1, X_2, ..., Xp$. The output values are normally denoted by *Y*. In general form this can be expressed as;

$$Y = f(x) + \varepsilon, \varepsilon \sim N(0, 1) \quad 3.13$$

f is a function of *X₁, X₂, ..., Xₚ* and ε is a random error term that is independent of *X* and has an average equal to zero. In formula 4.13, *f* is the systematic information *X* provides about *Y*. Statistical learning can thus mainly be said to refer to a set of approaches for estimating *f* (James et al., 2013). There are two reasons for estimating *f*, prediction and inference.

Predictors are in many situations easily accessible, while the response values cannot be easily obtained. In such settings, since the error term has an average of zero, we can predict Y, as expressed in equation 4.14.

$$\hat{y} = \hat{f}(X) \quad 3.14$$

Where $\hat{f}$ represents the estimate of *f* and $\hat{y}$ represents the resulting prediction of *Y*. The accuracy of $\hat{y}$ as a prediction for Y, depends on two factors, reducible error, and non-reducible error. Reducible error occurs due to inaccuracies related to the fact that $\hat{f}$ will generally not be a perfect estimate for *f*. This error can be reduced using the most appropriate statistical learning method. Non-reducible error will still exist, as even a perfect estimate for *f* will still be associated with error, giving the estimate the form, $\hat{y} = \hat{f}(X)$ as a result of the error term ε. This term cannot be predicted using X, so the error given by ε is not reduced no matter how good the estimate for f is (James et al., 2013). The reason why the error term ε is non-zero is because this term may contain information or variables that are useful for explaining Y, or because the clause contains information that cannot be measured. The errors in the prediction can be given by:

$$E(y - \hat{y})^2 = E\left[f(X) + \varepsilon - E(\hat{f}\right]^2 = \left[f(X) - E(\hat{f}\right]^2 + Var(\varepsilon) \quad 3.15$$

Where $E(y - \hat{y})$ represents the expected value of the squared differences between predicted and actual *Y*, and $Var(\varepsilon)$ represents the variance associated with ε. The first part of the equation, $\left[f(X) - E(\hat{f}\right]^2$, can be reduced while the last part, $Var(\varepsilon)$, cannot. The focus in statistical learning is to use techniques to estimate *f* so that the reducible error is minimized, to be able to predict *Y* as accurately as possible (James et al., 2017).

Inference is used when the goal is not to predict *Y*, but to estimate *f,* namely how *Y* changes as a function of the *X* variables, X_1, ..., X_p. The form of *f* is of great importance for understanding the movements in Y, as opposed to prediction, where the form of *f* is not as important if it gives accurate predictions. In such a setting, the purpose may be to understand which variables affects Y, what is the relationship between each variable and Y, and whether the relationship can be formulated linearly, or whether it is more complex (James et al., 2013).

3.6.2 Method Selection

Machine learning tools can be classified as either supervised or unsupervised.
Supervised learning is statistical models built for predicting or estimating output values based on one or more input values. Meaning that supervised learning has a starting value that guides the learning process. That is for each observation of the predictor's value $x_i, i = 1, \dots, n$ there is an associated response value y_i. With supervised learning, we want to adapt a model that relates the response value to the prediction value (James et al., 2013).
Unsupervised learning, however, have access to input values, but there are no output values that guide the result. That is, only if the input values are observed, and no measure for the output value. We have that for each vector i = 1, ..., n, we observe a vector of targets, but no corresponding response y_i. The models in unsupervised learning try to recognize patterns in the data, and there is often talk of clustering for non-supervised machine learning. For example, grouping companies based on accounting numbers. It is not possible to adapt a linear regression model to this data (Hinton & Sejnowski, 1999).

This study is based on supervised learning and can be continued as regression or classification problem. The classification for supervised machine learning is based on the response variables. Problems related to a quantitative response are often referred to as regression problems. Quantitative variables take numerical values, in our case stock returns. Classification problems, on the other hand, are problems that have qualitative response variables. The qualitative variables take values in one of K different classes or categories, such as "up" and "down" related to stock prices (Mohri et al., 2012). Regarding this thesis topic and issue, a classification problem using qualitative variables is preferred as this have shown better results compared to regression problems concerning financial issues (Enke, 2005; Leung et al., 2000)

There are multiple machine learning methods for supervised learning and choosing the right one for a given task, is an important part of the process. The goal is to choose the method that gives the best results for the selected data. To measure how well a model performs, one must measure how well the predictions match the observed data. In connection with measurement and model adaptation, the terms training and test data are introduced. Training data is observational data used during the training of models, while test data consists of the observations that are kept out when adapting models. Test data is used to examine how models perform when used on unknown data. For models, there is an expression of accuracy both regarding training data and test data. The expression for both training and testing accuracy when using qualitative variables is given below.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \qquad 3.16$$

$\hat{y}_i$ is the predicted class for observation number i using $\hat{f}$ and $I(y_0 \neq \hat{y}_0)$ is an indicator variable that is equal to 1 if $I(y_0 \neq \hat{y}_0)$ and 0 if $I(y_0 = \hat{y}_0)$. A value of 0 indicates a correct classifier and a value of 1 indicates a misclassification. This equation calculates the proportion of misclassifications and is referred to as the error rate of the training observations. This equation only calculates based on data used for model training and does not indicate how well the trained model fits on unknown data. This is tested by calculating the error rate for test observations given by the form (x_0, y_0), and is given by the equation 4.17, where $\hat{y}_0$ is the predicted class using $\hat{f}$ on the unused predictor x_0.

$$Ave(I(y_0 \neq \hat{y}_0)) \qquad 3.17$$

A good way to evaluate model fit is with skewness and variance. Skewness refers to the error that occurs by trying to make an approach to a complicated real problem using a less complex model. And Variance refers to how much $\hat{f}$ changes if estimated using a different data set for the training.

The machine learning model chosen for this thesis is Random Forest. This method is presented in the next chapter.

3.6.3 Random Forest

The "random decision forest" algorithm was first presented by Ho (Kam, 1995), who later wrote several articles on the method "the random subspace" which performs random selection from a selection of factors used to grow decision trees (Hastie et al., 1998). An extension of this was presented by Breiman (Breiman, 1996, 2001), also inspired by Amit and Geman (1997) who introduced the idea of searching through a random selection of available decisions by splitting decision trees. In this extension, it appears that Random Forest uses bagging to combine a group of decision trees to reduce variance and the effect of noise (Breiman, 2001). This means that a combination of decision threes and bagging is the foundation for Random Forest.

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features (Breiman, 2001). This method works in many ways like human behavior and has the advantage of being a straightforward way of illustrating a decision process. Figure 9 shows a simple decision three on how a soldier would decide on what to do in a combat situation. This principle can be transferred to a variety of problems.

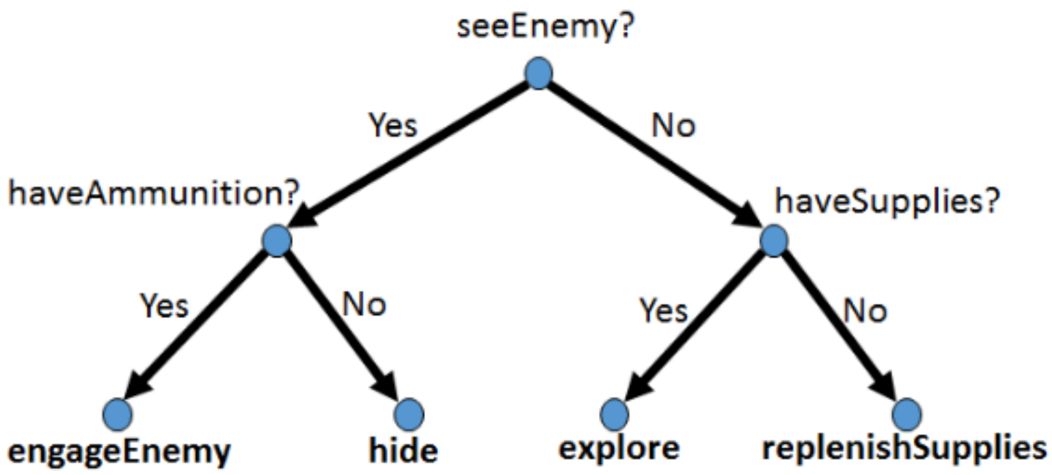

Figure 9: Simple Decision Three (Glavin, 2015)

Bagging is an abbreviation for bootstrap aggregation, as the different versions are formed by creating bootstrap replications of the training set and using these as new training sets. The bootstrapping method was first introduced by Efron (1979) and further developed in a later publication (1987). The idea behind the method is that a sample of a population contains all available information about complete population. Furthermore, repeated sampling with withdrawal will be the best approach for making repeated selections from the sample population. This is illustrated in Figure 10.

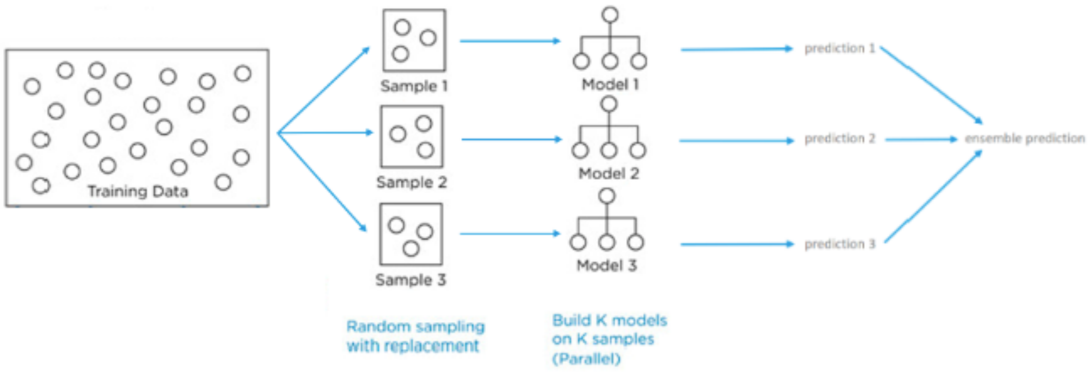

Figure 10: Bootstrap Aggregation

Through the decoration of the trees, the Random Forest method is an improvement over traditional bagging. As in bagging, the model is built by generating a number of decision trees on bootstrapped training selections. The difference is when these decision trees are built, a random selection of predictors will be selected as split candidates from the full range of predictors each time splitting is considered. Furthermore, the split can only use one of these predictors. A new selection of m predictors is thus selected at each split, and m is typically selected so that $m \approx \sqrt{p}$. Thus, the number of predictors at each split is approximately equal to the square root of the total number of predictors, *p.* The difference between bagging and Random Forest is the choice of size for the sample predictors, *m*. This means that if a Random Forest is built with *m* = *p*, the result will be the same as for bagging. The algorithm is illustrated in figure 11.

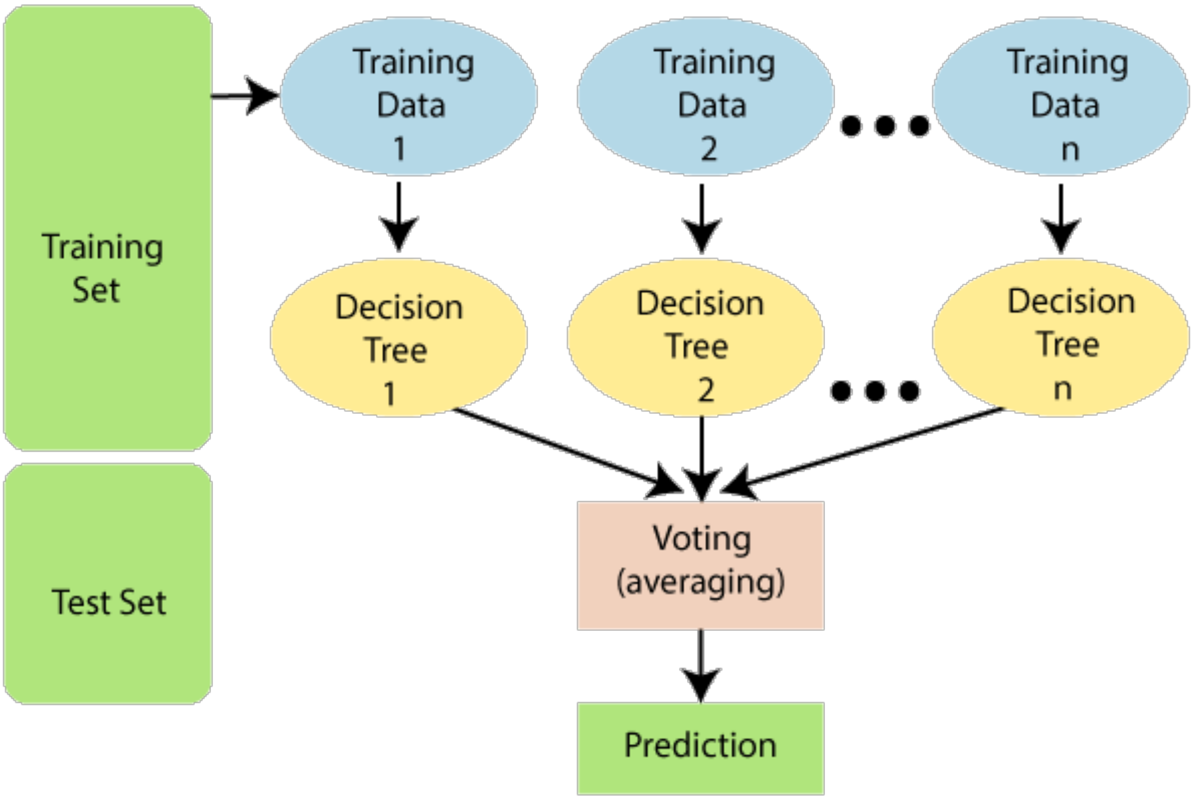

Figure 11: Random Forest Algorithm

To decide on which available predictor that should be used in the split, the split criterion *Gini Impurity* is introduced. This is a measure of the probability of an incorrect classification if a random observation in the dataset is classified according to the dataset's class distribution. The expression for this is given by:

$$I_G(k) = \sum_{i=1}^{C} P(i) * \big(1 - p(i)\big) \; = 1 - \sum_{i=1}^{C} (P_i)^2 \qquad 3.18$$

The predictors, or independent variables for this Random Forest model consists of a selection of different technical indicators. Research have shown that technical analysis is widely used by portfolio managers around the world and have shown promising results (Menkhoff, 2010).

3.6.4 Technical Analysis

Technical analysis is a theory which claims that a security's past activity and change in price can be good indicators of future price movement. this type of analyses consists of a wide variety of different technical indicators, which is used to predict future trends and sentiments for future price development of a security. This theory has its origins from the Dow Theory based on a series of articles by Charles Dow, which has been further developed by Hamilton and Rhea (Rhea, 2013). The theory is based on the assumptions that Markets effectively show the values that represent factors which affect the price of a security; however, market price movements are not purely random. This means that this theory contradicts the market efficiency theory, calming that historic price cannot be used for making predictions about future price development.

The basic precondition for technical analysis is that all known fundamental factors are included in the price, and that there is no need to emphasize these. In this way, technical analysts do not try to measure the fundamental value of a security, but rather try to identify patterns and trends that indicate what the security will do in the future (Bodie et al., 2018).

3.7 Constructing the Random Forest Portfolio

The portfolio constructed based on machine learning, uses the Random Forest algorithm to predict the probability of the asset rising in price over the next period. The calculation is done on daily time series observations, and follows the original Random Forest model given by Leo Breiman (Breiman, 2001). As mentioned, the technical execution of the model is done in the open-source program R, with the use of the Random Forest, Ranger and caret package.

3.7.1 Response Variable

To construct a response variable for a Random Forest algorithm based on time series data, we must convert the response variable to a classification predictive modeling. Research has shown that a classification problem yields better results when predicting on data from the financial market (Novaković et al., 2017). In this case the change in closing price from one day to another is either classified as “up” or “down”, meaning that the closing price has either risen or fallen in price over the previous period. This is expressed by constructing a trend indicator and is the response variable for this Random Forest model. This is expressed in the formula below.

$$Trend = if[(close\ prise_n - close\ price_{n-1}) > 0] = up \qquad 3.19$$

$$Trend = if[(close\ prise_n - close\ price_{n-1}) < 0] = down \qquad 3.20$$

3.7.2 Independent Variables (Predictors)

The technical indicators included in this analysis is listed below. These indicators are constructed in R using the TTR and Quantmod package.

1. MACD: Moving Average Convergence/Divergence

   This indicator consists of two different moving averages, comprised of a MACD line and a signal line. A high MACD number indicates the probability of an upward trend in the future, whereas a negative MACD value indicates the opposite. If the MACD line - Signal line crosses zero, the MACD line's trend is likely to alter. This is sometimes referred to as a MACD oscillator (Lo et al., 2000). This indicator is expressed by two equations, where equation 4.21 indicates the MACD line, and equation 4.22 indicates the signal line.

$$MACD = EMA_{12}(P) - EMA_{26}(P) \qquad 3.21$$

$$Signal = EMA_9(MACD) \qquad 3.22$$

2. ROC: Rate of Change

   Using rate of change as a measure, you can determine the percentage change in value over a specified period. ROC is calculated by dividing the current close price of a stock by the close price from an earlier period (Lo et al., 2000). In this case a period is 14 days. This indicator is a momentum indicator, and is calculated as follow:

$$ROC = \left(\frac{close\ price_t}{close\ price_{t-14}} - 1\right) * 100 \qquad 3.23$$

3. STO: Stochastic

Stochastic oscillators are momentum indicators that compare a particular closing price of a security to a range of its prices over time. Taking a moving average of that result or adjusting the time period reduces the sensitivity of the oscillator to market movements (Lo et al., 2000). STO is expressed by dividing the last close price minus the lowest closing price in lookback window with the highest closing price in the lookback window with the lowest closing price. The lookback window is 14 days, giving equation 4.24.

$$STO = \left(\frac{close\ price - \ low_n}{high_n - low_n}\right) * 100 \qquad 3.24$$

4. ATR: Average True Range

In technical analysis, the average true range (ATR) is an indicator of market volatility. Typically, it is derived from a series of 14-day simple moving averages. Initially developed to be used in commodities markets, it has been adapted for use in all types of securities (Lo et al., 2000). The first step is to calculate the true range. This is the maximal price range for a given day during the period and the absolute value for the highest price minus current closing price and the lowest closing price minus the current closing price. This is expressed in equation 4.25. Then the sum of all true ranges is summed and divided on the time period used for the calculations, which in this case is 14 days. This gives the average true range and is expressed below:

$$TR = Max[(H - L), Abs(H - close\ price), Abs(L - close\ price)] \qquad 3.25$$

$$ATR = \left(\frac{1}{n}\right) \sum_{(i=1)}^{(n)} TR_i \qquad 3.26$$

5. RSI: Relative Strength Index

One of the most popular and widely used momentum oscillators is the Relative Strength Index (RSI). It is a measure of both the speed and the rate of change in price movements within the market (Lo et al., 2000). This is calculated by dividing number of days with closing price higher than opening price divided with number of days with closing price lower than opening price. This is expressed as follows:

$$RSI = 100 - \left(\frac{100}{1 + \frac{n_{up}}{n_{down}}} \right) \qquad 3.27$$

3.7.3 Predicting the Portfolio

For each trading period the probability for each asset is predicted, giving the probability of a stock going up in price over the next period. Since daily data is used, we convert this to monthly probabilities. This is expressed as follow $\widehat{Prob_{t+m}^{s}}$. All assets with >50% probability of rising in price over the next period, is included in the portfolio for one period and is rebalanced every period. If there are no assets with correct probability cutoff, the portfolio will consist of cash only.

4 Results

In this section we will present the descriptive statistics for each portfolio and their result for the six-year test period. At the end of the chapter, we will present our findings from the regression analysis and explain how the factors in the model explain variance in the portfolios.

The performance for each of the five portfolios is presented in figure 12 and show that the performance for all portfolios except the Random Forest portfolio is fairly equal, with a total return between 91,5% - 98,9%, all outperforming the S&P 500. The Random Forest portfolio has performed significantly better, giving a total return of 133%. This shows that using the Random Forest method to optimize portfolios yields the highest return. Another interesting point is that when all portfolios get a bigger pullback, the Random Forest portfolio seems to have a significantly smaller pullback. It seems that the Random Forest algorithm manage to predict when the market performs poorly and stays out of the market at these times. This is clearly visible in the early spring of 2020, when the financial market got hit by the COVID-19 pandemic, dragging the S&P 500 down 29% at most. Our portfolios, except Random Forest, all fall accordingly, the minimum variance portfolio losing the hardest as it had the highest return going into 2020. The buy and hold portfolio follow the naive 1/N portfolio performing the weakest of all portfolios throughout our testing period strengthening the findings of Hillard & Hillard (2015), although the minimum variance portfolio closes just 1% above, 1/N.

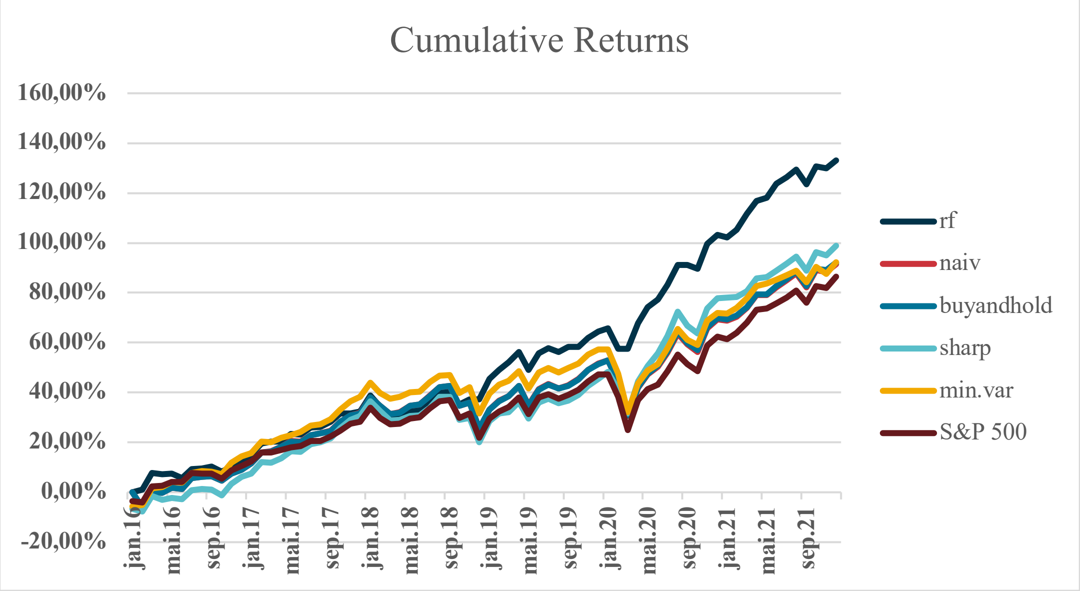

Figure 12: Cumulative Returns for the Five Portfolios During Test Period

4.1 Summary Statistics

Table 4 shows summery statistics for all portfolios and gives a quick overview of all portfolios. A normal assumption in finance is that greater returns relate to more risk, and a portfolio with high returns usually have a greater standard deviation. This does not seem to be the case for the findings in this thesis. If we examine the table, we get the indication that RF is performing better, and gives a greater return for a lover risk. We will analyze this in greater detail.

Table 4: Summary Statistics Table for All Portfolios with Monthly Returns.

	RF	***1/N***	***B&H***	***SR***	***MV***
Mean	**0,0185**	**0,0127**	**0,0128**	**0,0137**	**0,0128**
Standard error	0,0042	0,0051	0,0052	0,0052	0,0053
Median	0,0193	0,0203	0,0202	0,0159	0,0179
Std	**0,0357**	**0,0437**	**0,0437**	**0,0445**	**0,0446**
Variance	0,0013	0,0019	0,0019	0,0020	0,0020
Kurtosis	**0,7848**	**2,0300**	**1,9392**	**0,7319**	**2,7656**
skewness	**-0,2659**	**-0,9865**	**-0,9602**	**-0,4010**	**-1,0945**
Minimum	**-8,17 %**	**-13,78 %**	**-13,60 %**	**-9,70 %**	**-15,36 %**
Maximum	**10,13 %**	**11,29 %**	**11,37 %**	**12,92 %**	**11,52 %**
Sum	**133,05 %**	**91,47 %**	**92,09 %**	**98,86 %**	**92,25 %**
Observations	72	72	72	72	72

To compare risk and return we use annual Sharpe-Ratio as it tells us how much return we got for each percent standard deviation. This metric gives a good indication of the risk/reward connection, and what portfolio that gives the best returns and lowest risk. The formula for the Sharpe-Ratio is presented earlier and shows that an increase in returns (numerator) or a decrease in Standard deviation (denominator) will give a higher Sharpe-Ratio.

As figure 13 shows, the RF portfolio has the highest Sharpe-Ratio, just under 2. All the other portfolios are just above 1 and does not give a good risk/reward ratio. Figure 14 is a decomposition of the Sharpe-Ratio and shows both the return which is the numerator and the standard deviation (Std) which is the denominator. This figure shows what driving forces behind the Sharpe-Ratio. It is very clear that the Random Forest portfolio stands out, with much greater return for a lower risk then the other portfolios.

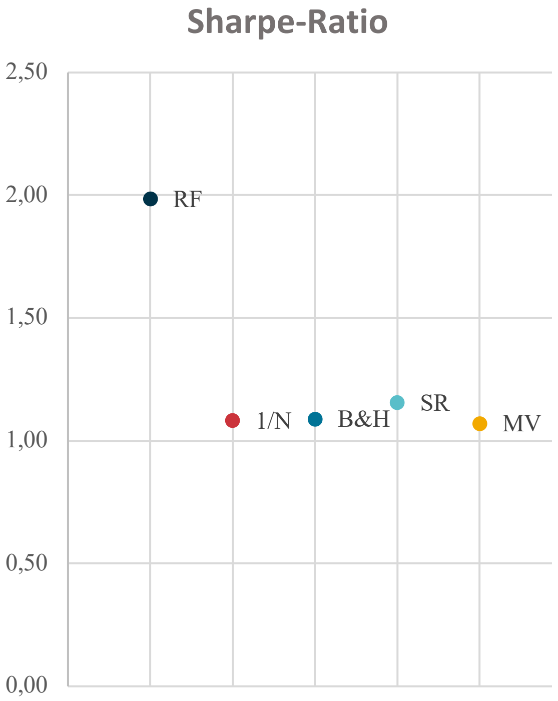

Figure 13: Sharpe-Ratio for All Portfolios

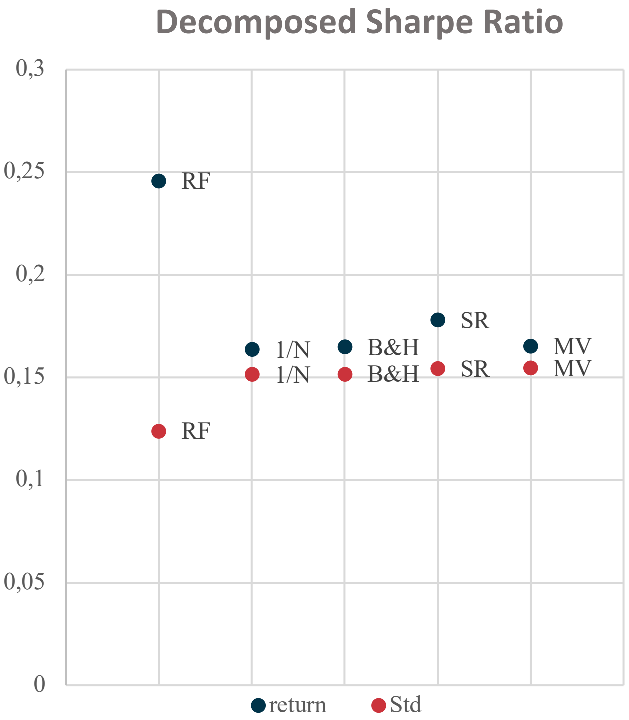

Figure 14: Decomposed Sharpe-Ratio for All Portfolios

4.2 Return Distribution

In this chapter we will compare returns and discuss some of the risk factors using normal distribution histograms and looking at the trend for the return distributions. Note that all distributions are constructed using our monthly returns for each portfolio, 72 returns in total during the test period of 6 years.

The naive portfolio is the lowest yielding portfolio and has returned 91,47% in total over our test period of 6 years. This portfolio has a skewness of -0,98 and a kurtosis of 2,76, meaning that the left tail is fatter than the right, the kurtosis also indicates that the distribution is not normally distributed. This means that the overall risk of the portfolio performing extreme negative value is greater than an extreme positive value. If we look at a histogram of the return distribution, we can visually see that although the portfolio mostly yields 0,15-2,93% monthly, the extreme values are dominantly left sided. If we look at the distribution trend as shown in figure 16, then this also shows that the most extreme values are in the negative direction. From the descriptive statistics we can see that max drawdown is -13,78% and max gain is 11,29%.

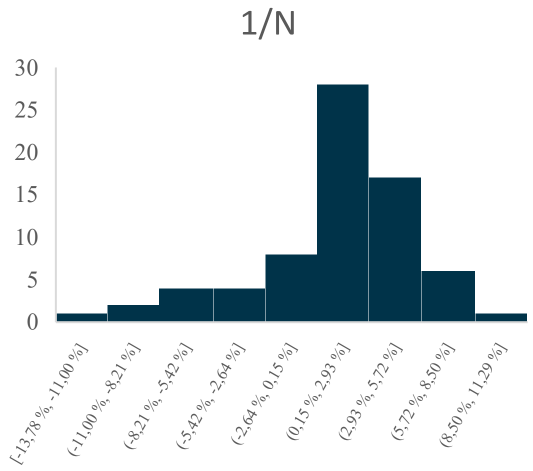

Figure 15: Return Distribution for 1/N.

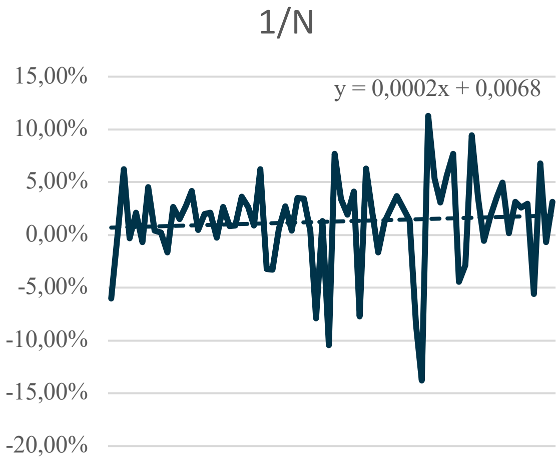

Figure 16: Return Distribution Trend 1/N

The next portfolio is the buy and hold portfolio, which has yielded a return of 92,09% over the whole period. This portfolio is negatively skewed -0,96 also with fatter tails on the left side. The kurtosis is 1,9392 meaning that the distribution is not normally distributed. This is similar to the naive portfolio with a more or less equal return. The max drawdown is -13,60% and maximum gain is 11,37% with a standard deviation of 0,0437, same as the naive. So, the risk is more or less equal for these two portfolios. This portfolio is not rebalanced each month, so all costs related to this are saved. This portfolio does therefore perform better in reality, but this study does not include transaction costs for simplicity reasons. The return distribution and return distribution trend shown in figures 17 and 18 indicate the same risk and return as for the naive portfolio. The similarities are expected due to the fact that the weights of these two portfolios are not that different, shown in figure 4 and 5. If our study had involved a few stocks of companies instead of ETFs the results would probably be much different, as one of the companies could have dictated the variance and return of the whole portfolio.

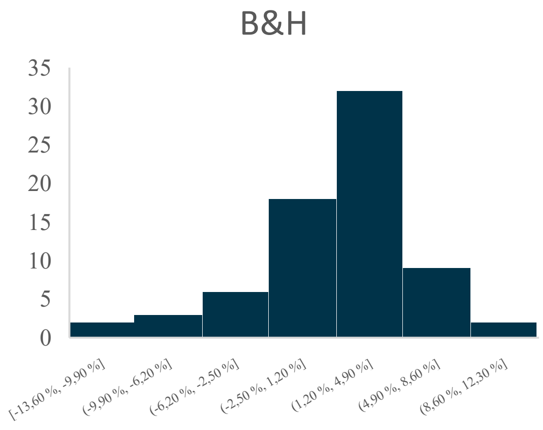

Figure 17: Return Distribution for B&H.

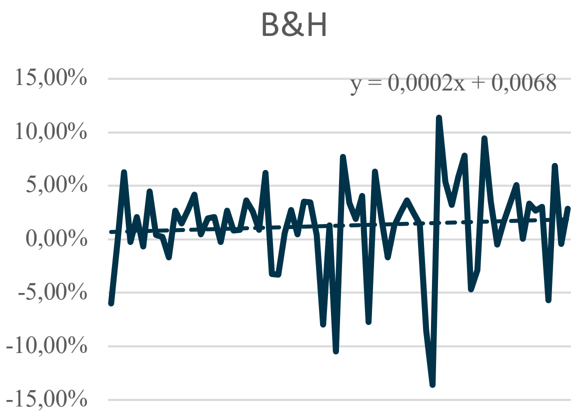

Figure 18: Return Distribution Trend for B&H.

The next portfolio is minimum variance portfolio, which has returned 92,25% over the total period. The skewness is -1,0945 indicating fatter tails on the left side. The kurtosis is 2,7656 meaning that this portfolio is not normally distributed and as shown in figure 19 we can see a even greater part of montly returns being in the -0,56-3,14% window. Max drawdown for this portfolio is also higher with -15,36% and max gain is surprisingly the same as the two previous on 11,52%. The standard diaviation is 0,0446, wich is the highest of all portfolios. This is interesting since the strategy is constructed with the weights that gave the lowest possible variance in return the previous month and indicates some sort of build-up effect in varaince. If we look at the return disribution and the the distribution trend for this portfolio, we can see the same tendency as previous portfolios, but with greater standard deviation.

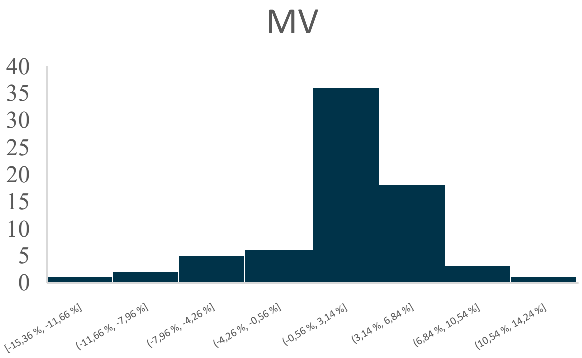

Figure 19: Return Distribution for MV

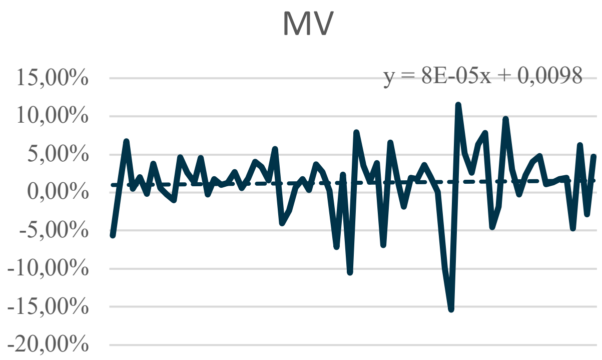

Figure 20: Return Distribution Trend for MV

The max Sharpe-Ratio portfolio has performed 98,86% over the period. The skewness is -0,4010 meaning that this portfolio is skewed to the left but is classified as normally distributed. The kurtosis of 0,7319 also indicates more of a normal distribution. The max drawdown for this portfolio is -9,7% and max gain is 12,92%. This is interesting for a portfolio trying to maximize return at any given point of risk. As mentioned above, we do not expect this portfolio to have a lower risk than the minimum variance portfolio, and although we expected a higher return and Sharpe-Ratio, it is interesting that the distribution of returns is so normally distributed. The return distribution trend, figure 21, indicates a higher standard deviation than the minimum variance at first glance, but they are in total almost similar at 4,46% and 4,47%.

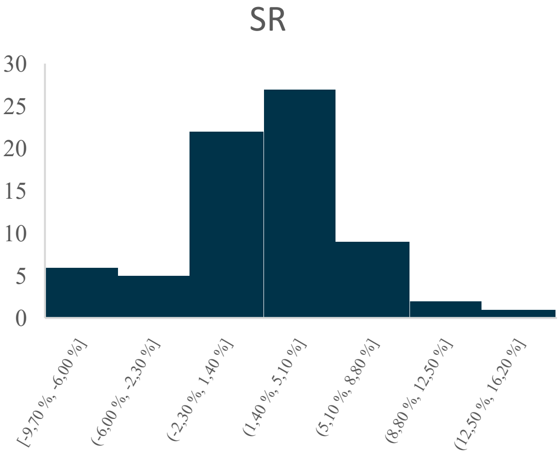

Figure 21:Return Distribution for SR

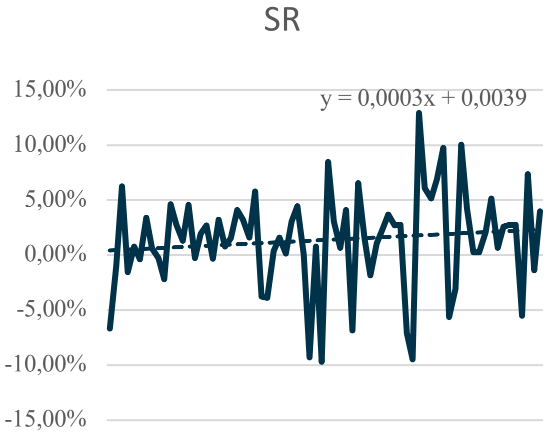

Figure 22: Return Distribution Trend for SR

The last and best performing portfolio is the Random Forest portfolio. This portfolio has a skewness of -0,2659 and a kurtosis of 0,7848, meaning that this portfolio is very close to normally distributed. The maximum drawdown for the portfolio is -8,17% and maximum gain is 10,13%. In other words it does not yield extreme returns, but at the same time avoid extreme losses. If we look at the return distribution and return distribution trend, this shows that the returns have lower standard diviation and is normaly distributed, with most returns being in the 1,40-5,10% monthly window. The biggest difference compared to previous portfolios, is that a lager number of returns are in the 4,03-8,10% window. This sums up to the same conclusion as mentioned earlier that the Random Forest porfolio neglects extreme losses and seems to avoid negative returns in general.

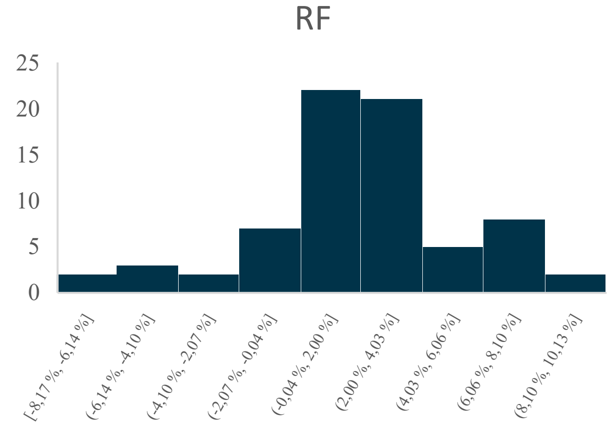

Figure 23: Return Distribution for RF

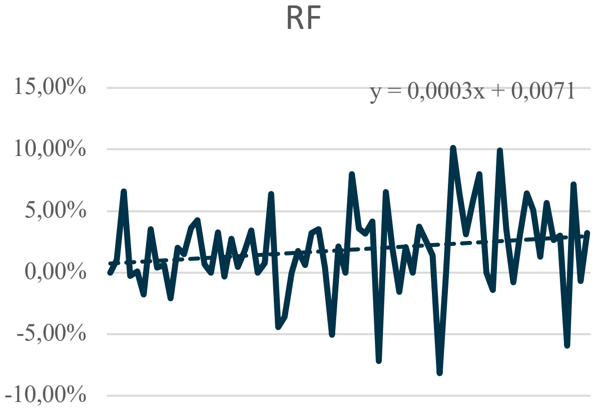

Figure 24: Return Distribution Trend for RF

4.3 Regression Results

Note: This table shows the results from 5 regression analysis. The dependent variables (portfolios) are listed on top with their coefficients on the left side. The intercepts are the unexplained factors and tells us how much of the variance in the portfolio cannot be explained by the factor's variance. Under each p-value are stars explaining the significance level for each factor. Three stars (***) means the factor is significant on a 0% level, meaning the p-value is between 0 and 0,001, two stars (**) means the p-value is between 0,001 and 0,01, one star (*) is between 0,01 and 0,05, period symbol means between 0,05 and 0,1 and lastly no symbol means the p-value is between 0,1 and 1.

Table 5: Results from Regression Analysis

	RF	1/N	B&H	MV	SR
Intercept (α)	0,007	-0,002	-0,002	-0,001	-0,002
p-value	0,001	0,013	0,014	0,080	0,201
	**	*	*		
Mkt	0,721	1,029	1,031	1,029	1,048
p-value	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16
	***	***	***	***	***
SMB	-0,303	-0,154	-0,150	-0,130	-0,109
p-value	0,001	0,000	0,000	0,000	0,021
	***	***	***	***	*
HML	-0,225	-0,025	-0,036	0,133	-0,088
p-value	0,001	0,328	0,165	0,000	0,018
	**			***	*
MOM	-0,131	0,058	0,061	0,077	0,102
p-value	0,064	0,035	0,027	0,006	0,010
	.	*	*	**	**
R²	0,788	0,979	0,979	0,979	0,958

Our result from the regression shows that the naive 1/N portfolio and B&H portfolio shows almost the exact same coefficients for all factors. There is a slight difference in SMB, but not much. The Mkt and SMB both have a significance level between 0 and 0,001. This means both are significant under the 95th percentile (α = 0,5), in other words, these two factors add value to our model and is explaining some of the variance in the chosen two portfolios. HML and MOM is not significant for these two portfolios, neither is the intercept.

The minimum variance portfolio (MV) and tangency portfolio (SR) both have the same coefficient for the market factor and SMB, although SMB is not significant for the SR portfolio. Same applies for the HML factor which is significant with a positive correlation of 0,133 for the MV portfolio but is negative and not significant for the SR portfolio. For the MV, SR, 1/N and B&H portfolio the R-squared is almost the same (0,958-0,979) which implies that the independent factors variance can explain over 95% of the individual portfolio's variance.

The most interesting find from this analysis is with the Random Forest portfolio. All the factors are significant except the MOM factor. This is interesting because the technical indicators in the Random Forest model mostly rely on price trends. This indicates that the model uses the technical indicators to break negative trending momentum, but still yields positive return with momentum. It is the only portfolio with a significant intercept (α) which means the model have a significant excess return of 0,7% monthly that cannot be explained by the 4-factor model. The R-squared is also lower for the Random Forest portfolio, meaning the model only explains 78,8% of the variance in the portfolio.

4.4 Random Forest Portfolio Findings

If we look at the performance for each portfolio, we can see that for the first three years that the return development was approximately equal between all portfolios. After this period ended in January of 2019, the Random Forest portfolio starts to outperform all the other portfolios. The driver behind this seems to be that all other portfolios get a pullback, while the Random Forest portfolio stays on the same trend. When the COVID-19 crash came in mars 2020 the same phenomena occur. All portfolios get a steep drop while the Random Forest portfolio only gets a small pullback and then stays out of the market. The Random Forest portfolio then had a head start when the market started to regain its losses. The other portfolios seem to have a better bounce back after COVID-19 while the Random Forest portfolio again shows a slow and steady return.

Based on these findings, the reason for the significantly better performance for the Random Forest portfolio seems to be the ability to predict when the market is getting a bigger pullback. As discussed in the return distribution chapter, the Random Forest portfolio does not have the most extreme distribution but has a stable and positive return from month to month.

We have a six-year testing period and during this period there is two occasions where the pullback is >15%, one in early 2019 and the other in mars 2020. Since these two incidents seems to have a significant impact on the overall performance for all portfolios, the testing period can be a driver behind the price development for this portfolio. If the testing period was longer and included more periods with pullback in the market, then we could with more certainty conclude that the phenomena is real and not just a coincidence. But nevertheless, the findings are very interesting and the potential to do more research based on our findings is present.

The Random Forest model uses technical indicators as input, which is based on historic prices. Using historic prices to predict future price development is a direct contradiction to the market efficiency theory, which clearly states that this is not possible. Our findings indicate that using historic prices can create excess returns, as mentioned before, the timespan for this thesis is too short to reliably challenge the theory of market efficiency.

5 Conclusion

To sum up our analysis and results we have similar findings as previous research listed in table 1. The only exclusion being that one of our optimalization by Markowitz weights, max Sharpe-Ratio, outperformed the two simple strategies, while the minimum variance did not. Controversially the minimum variance portfolio had the highest risk of all portfolios with a standard deviation of 15,44% annually and although it yielded slightly higher return than the two simple strategies, the Sharpe-Ratio was a great deal lower. This indicates some sort of a reversed effect when the portfolio had the lowest risk the month before, the coming month seems to be riskier than any of the other portfolios.

The two simple strategies Buy & Hold and Naive 1/N portfolio had almost identical results and with a significant market factor and a R-squared of 97,8%, these two portfolios indicate mostly a good period for the chosen ETFs. None of our first 4 portfolios had a statistically significant intercept (α) meaning we did not find a significant excess return in the portfolio. For these portfolios the Market Efficiency Theorem (MET) seems to hold, and we are left with the indication of an efficient market. Since we are weighting ETFs and no single companies, the weights of the buy and hold portfolio does not get over weighted by one asset, hence the similar results as the naive 1/N portfolio.

The most complicated strategy tested, the Random Forest portfolio, had the highest return, lowest risk and a very interesting normal distribution. The number of negative returns were greatly outnumbered by the positive observations. Even though there were a few extreme observations, it was few compared to the other strategies. We managed to find a positive significant intercept (α) from the regression analysis, yielding 9% annually. The Fama French Carhart 4-factor model could explain 78,8% of the variance in the portfolio, while it explained 95,5-98,8% of the other portfolios. Even though this return is extremely high compared to the other portfolios, it is not as high as Tan et al. (2019) and Kilskar (2019). This is most likely because we use a composition of defensive assets (ETF`s) to make up our portfolios, and thus yields lower returns overall. Both previous studies found that the Random Forest portfolio had declining results during their test period. This is not the case for our study as we can see from figure 12, the portfolio performs strongest towards the end of the testing period. As discussed earlier, we believe this is caused mainly by the fact that the

Random Forest portfolio does not pull back as much as the other portfolios, and thus gets a head start when the market is turning bullish.

Based on our findings we can conclude that the problem statement for this thesis indicates that an advanced form of portfolio allocation yields better results. The Random Forest portfolio yields 9% annual excess returns which could not be explained by the Fama French Carhart 4-factors. This is not the case for the other portfolios, these factors explain 98% of the variance for these portfolios, and does not yield excess returns.

5.1 Weakness With Our Study

To simplify this study, we have implemented some limitations that are necessary to mention.

1. The testing period is limited to six years. This might impact the reliability of the study since some important findings only occurs a few times and the market was generally more volatile compared to the timeframe in previous studies.
2. We have not included transaction cost to our study so our comparing of the buy and hold portfolio to portfolios with frequent rebalancing is not reliable in practice.
3. We did not include risk free rate, but as the risk-free rate have been so low during our period, we do not think it would affect our results in a significant way.
4. We have not included shorting in our study.
5. We have chosen to only look at the combination of five different ETFs, all from the US market.

5.2 Further Research Topics

Based on our findings, there is multiple angles for further research. The most interesting finding we have found in this paper, is the fact that the RF portfolio gets significantly lower pullbacks compared the other portfolios. Extending the testing period to see if the phenomena is consistent over a wider timeframe is a good basis for further research. Another topic is to understand what the driving factors behind the performance of the Random Forest portfolio is, and how it manages to stay out of the market under big pullbacks. Including assets from other markets could also be done, to confirm that the findings is reliable and valid.

References

Albrigtsen, B. (2007). *Effekten av endringer i lakseprisen på aksjekursen til noen utvalgte lakseselskaper på Oslo Børs* Universitetet i Tromsø.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, *9*(7), 1545-1588.

Amundsen, H. T., & Statistisk, s. (1980). *Lineær regresjon uten konstantledd : litt elementær informasjon* (Vol. 80/32). SSB.

Bjordal, A., & Opdahl, E. (2017). *Portfolio optimization in the cryptocurrency market : an evaluation of the performance of momentum strategies in the cryptocurrency market and cryptocurrency's place in an optimized investment portfolio* Norges Handelshøgskole.

BlackRock, I. (2022). *iShares MSCI USA ESG Select ETF*. BlackRock, Inc. Retrieved 02.05 from https://www.ishares.com/us/products/239692/ishares-msci-usa-esg-select-etf

Bodie, Z., Kane, A., & Marcus, A. J. (2009). Investments, 8. Baskı. In: Mc Graw Hill Yayınevi, Singapur.

Bodie, Z., Kane, A., & Marcus, A. J. (2018). *Investments* (11th ed.). McGraw-Hill.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine learning*, *45*(1), 5-32.

Brinson, G. P., Hood, L. R., & Beebower, G. L. (1986). Determinants of Portfolio Performance. *Financial analysts journal*, *42*(4), 39-44. https://doi.org/10.2469/faj.v42.n4.39

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance (New York)*, *52*(1), 57-82. https://doi.org/10.2307/2329556

Chan, N. H., & Wong, H. Y. (2013). *Handbook of financial risk management: simulations and case studies* (Vol. 12). John Wiley & Sons.

Cohen, G., & Cabiri, E. (2015). Can technical oscillators outperform the buy and hold strategy? *Applied Economics*, *47*(30), 3189-3197. https://doi.org/10.1080/00036846.2015.1013609

CorporateFinanceInstitute. (2015). *Regression Analysis*. CFI Education Inc. Retrieved 09.05 from https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy? *The Review of financial studies*, *22*(5), 1915-1953. https://doi.org/10.1093/rfs/hhm075

Doroghazi, R. M. (2020). The Myth of Buy and Hold Forever. *The American Journal of Cardiology*, *132*, 158.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, *21*(4), 460-480.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, *82*(397), 171-185.

Enke, D. (2005). Expert Systems with Applications. *Enke Thawornwong*, *29*, 927-940.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383-417. https://doi.org/10.2307/2325486

Fama, E. F. (1970). Session topic: stock market price behavior. *The Journal of Finance*, *25*(2), 383-417.

Fama, E. F. (1991). Efficient Capital Markets: II. *The Journal of finance (New York)*, *46*(5), 1575-1617.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56. https://doi.org/10.1016/0304-405X(93)90023-5 (Journal of Financial Economics)

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, *51*(1), 55-84.

Finance, B. D. (2017). *Minimum Variance Portfolio*. Breaking Down Finance. Retrieved 11.05 from https://breakingdownfinance.com/finance-topics/modern-portfolio-theory/minimum-variance-portfolio/

French, K. R. (2022). *Kenneth R. French*. Kenneth R. French. Retrieved 07.05 from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, *38*(8), 10389-10397. https://doi.org/10.1016/j.eswa.2011.02.068

Hastie, T., Tibshirani, R., & Friedman, J. (1998). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: SpringerVerlag, 2009*.

Hayes, A. (2021). *Fama and French Three Factor Model*. Dotdash Meredith. Retrieved 06.05 from https://www.investopedia.com/terms/f/famaandfrenchthreefactormodel.asp

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251.

Hilliard, J. E., & Hilliard, J. (2015). A comparison of rebalanced and buy and hold portfolios: does monetary policy matter? *Review of Pacific basin financial markets and policies*, *18*(01), 1550006.

Hinton, G., & Sejnowski, T. J. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Invesco. (2022). *Invesco QQQ ETF (QQQ)*. Invesco Ltd. Retrieved 02.05 from https://www.invesco.com/qqq-etf/en/home.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kam, H. T. (1995). Random decision forest. Proceedings of the 3rd international conference on document analysis and recognition,

Kendall, M. G., & Hill, A. B. (1953). The Analysis of Economic Time-Series-Part I: Prices. *Journal of the Royal Statistical Society. Series A (General)*, *116*(1), 11-34. https://doi.org/10.2307/2980947

Kosev, M., & Williams, T. (2011). Exchange-traded Funds. *RBA Bulletin (Print copy discontinued)*, 51-60. https://EconPapers.repec.org/RePEc:rba:rbabul:mar2011-08

Lettau, M., & Madhavan, A. (2018). Exchange-Traded Funds 101 for Economists. *Journal of Economic Perspectives*, *32*(1), 135-154. https://doi.org/10.1257/jep.32.1.135

Leung, M. T., Daouk, H., & Chen, A.-S. (2000). Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of forecasting*, *16*(2), 173-190.

Lintner, J. (1965). Security Prices, Risk, and Maximal Gains From Diversification. *The Journal of Finance*, *20*(4), 587-615. https://doi.org/10.2307/2977249

Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, *55*(4), 1705-1765.

Markowitz, H. M. (1952). Portfolio Selection. *Journal of Finance*, *7*, 77-91.

McDonald, J. H. (2009). *Handbook of biological statistics* (Vol. 2). sparky house publishing Baltimore, MD.

Menkhoff, L. (2010). The use of technical analysis by fund managers: International evidence. *Journal of banking & finance*, *34*(11), 2573-2586. https://doi.org/10.1016/j.jbankfin.2010.04.014 (Journal of Banking & Finance)

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning.[Sl]. In: The MIT Press.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, 768-783.

Nickolas, S. (2022). *SPY: SPDR S&P 500 ETF Trust*. Dotdash Meredith. Retrieved 02.05 from https://www.investopedia.com/articles/investing/122215/spy-spdr-sp-500-trust-etf.asp

Norrestad, F. (2022). *Number of ETFs globally 2003-2020*. statista.com. Retrieved 01.03.2022 from https://www.statista.com/statistics/278249/global-number-of-etfs/#statisticContainer

Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Milica, T. (2017). Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1), 39–46-39–46.

Patel, K. (2018). Demystifying Buffett's Investment Success. *Financial analysts journal*, *74*(4), 25-27.

Qian, E. E. (2014). To rebalance or not to rebalance: A statistical comparison of terminal wealth of fixed-weight and buy-and-hold portfolios. *Available at SSRN 2402679*.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing (vol 2016, 67, 2016). *EURASIP journal on advances in signal processing*. https://doi.org/10.1186/s13634-016-0382-7

Rhea, R. (2013). *The Dow Theory*. Barron's Educational Series.

Sharpe, W. F. (1964). CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK. *The Journal of finance (New York)*, *19*(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Shilling, A. G. (1992). Market Timing: Better Than a Buy-and-Hold Strategy. *Financial analysts journal*, *48*(2), 46-50. https://doi.org/10.2469/faj.v48.n2.46

Shukla, R., & Bogle, J. C. (1994). Bogle on Mutual Funds: New Perspectives for the Intelligent Investor. In (Vol. 49, pp. 750-754). Cambridge: Cambridge: American Finance Association.

Sotiropoulos, D. P., & Rutterford, J. (2019). Financial diversification strategies before World War I: Buy-and-hold versus naïve portfolio selection. *Business history*, *61*(7), 1175-1198. https://doi.org/10.1080/00076791.2018.1512097

SSGA. (2022). *SPDR® Dow Jones® Industrial Average ETF Trust*. State Street Corporation. Retrieved 02.05 from https://www.ssga.com/us/en/intermediary/etfs/funds/spdr-dow-jones-industrial-average-etf-trust-dia

Sydsæter, K., & Hammond, P. J. (2008). *Essential mathematics for economic analysis*. Pearson Education.

Tan, Z., Yan, Z., & Zhu, G. (2019). Stock selection with Random Forest: An exploitation of excess return in the Chinese stock market. *Heliyon*, *5*(8), e02310-e02310. https://doi.org/10.1016/j.heliyon.2019.e02310

Tang, G. Y. N. (2004). How efficient is naive portfolio diversification? an educational note. *Omega (Oxford)*, *32*(2), 155-160. https://doi.org/10.1016/j.omega.2003.10.002 (Omega)

The Vanguard Group, I. (2022). *VUG Growth ETF*. The Vanguard Group, Inc. Retrieved 02.05 from https://investor.vanguard.com/etf/profile/VUG

Utami, W., & Nugroho, L. (2017). Fundamental versus technical analysis of investment: Case study of investors decision in Indonesia stock exchange. *The Journal of Internet Banking and Commerce*, 1-18.

Wessel, E. M. D. (2021). *What did the Fed do in response to the COVID-19 crisis?* The Brookings Institution. Retrieved 18.05 from https://www.brookings.edu/research/fed-response-to-covid19/

Zhu, M., Philpotts, D., & Stevenson, M. J. (2012). The benefits of tree-based models for stock selection. *Journal of Asset Management*, *13*(6), 437-448. https://EconPapers.repec.org/RePEc:pal:assmgt:v:13:y:2012:i:6:d:10.1057_jam.2012.17