Classifying High-Growth Manufacturing Firms on the Swedish Stock Market

A Comparative Study Between the Logistic Regression, Support Vector Machine and Artificial Neural Network

SCHOOL OF ECONOMICS AND MANAGEMENT
Lund University

William Fridström

Supervisor: Luca Margaritella

Department of Economics

Master Thesis II, May 2023

ABSTRACT

This is a comparative study between two modern machine learning algorithms, the Support Vector Machine and Artificial neural network, and one traditional econometric model, the Logistic regression. The main objective is to compare their performance by classifying high-growth companies. The study uses panel data from 156 manufacturing firms on the Swedish stock market between 2018 and 2021. The results show that in line with previous studies, the modern machine learning algorithms, Support Vector Machines and Artificial neural networks perform better in classification accuracy and misclassification rate compared to the logistic regression when classifying which manufacturing firms were high-growth during the four years studied. This study recommends adopting the Support Vector Machine algorithms for high-growth classification modelling.

CONTENTS

1 INTRODUCTION

Entrepreneurs, policymakers, and investors are highly interested in anticipating which companies will experience high growth (Henrekson & Johansson, 2010; Mason & Brown, 2013). High-growth firms have garnered much attention since David Birch's pioneering statistical study in 1979 that sheds light on the vast impact these high-growth firms has on the economy, owing to their remarkable ability to generate employment, wealth, and significant contributions to innovation and productivity growth. A better capacity to forecast high-growth firms is essential for investors looking to direct investments towards promising firms, policymakers seeking to develop an effective framework to promote job creation, and entrepreneurs with aspirations for growth.

However, as many previous researchers have concluded, the ability to forecast what firms will experience high growth is limited, "The most elementary 'fact' about corporate growth thrown up by econometric work on both large and small firms is that firm size follows a random walk" (Geroski, 2000: p. 169). If this statement holds, creating a model to classify these high-growth firms accurately should be challenging.

There are numerous methods available to gauge the size and growth of firms. Assets, number of employees, value-added, profits, and revenue are all recognised approaches for assessing the size of an enterprise. Similarly, growth can be measured using various techniques, including relative and absolute growth, as well as unorthodox methods such as the "Birch index," which is calculated by sorting a sample of businesses based on their annual employment growth rate over a specific period, typically one year. The firms are then ranked in descending order according to their growth rates. The index is created by summing up the cumulative percentage of employment growth contributed by each firm as one moves down the ranked list. (Delmar et al., 2003).

This study aims to employ and compare the performances of three classification models: logistic regression, support vector machine (SVM), and artificial neural network (ANN). The object of the models will be classifying if a publicly traded Swedish manufacturing firm has experienced a high-growth rate, defined as a growth rate of 20% or above according to the OECED (2021), between 2018 and 2021. A panel of 156 manufacturing companies traded on

the Swedish stock market is considered. Dimension reduction is employed to combat overfitting due to noise and multicollinearity using Principal component analysis (PCA). The models are compared based on four measures, Accuracy, Misclassification rate, Specificity, Sensitivity and AUC-score, focusing on performance on test data. This study employs revenue and relative growth as firm size and growth measures.

Numerous studies have been conducted to classify high-growth companies using logistic regression and artificial neural networks. For instance, Has et al. (2016) find that the more modern machine learning algorithms outperform the standard logistic regression regarding classification accuracy when analysing Croatian businesses over three years. They measure growth in terms of assets and use companies from various sectors to conduct the analysis. Similarly, Zhou & Gumbo (2021) conduct a similar study, incorporating another machine learning model, the support vector machine, to classify whether a firm had experienced growth during one year. They use small and medium-sized firms from the manufacturing industry in South Africa. The same conclusion is drawn, demonstrating that the classification abilities of the machine learning models (SVM and ANN) outperformed that of logistic regression.

This study aims to classify whether a company is "high-growth" during a four-year period using revenue growth instead of asset growth, as used in Has et al. (2016). The reasoning for this choice instead of employment or asset growth is, as according to Hermelo & Vassolo (2007), there is a challenge when measuring growth in terms of employment because it tends to be biased against capital-intensive firms. On the other hand, using assets as a measure of growth rate discriminates against labour-intensive firms.

The manufacturing industry in Sweden will be analysed, specifically publicly traded manufacturing companies on the Swedish stock market (NASDAQ). The choice of the manufacturing industry is justified for two reasons: First, revenue is a reliable measure of growth in the manufacturing industry since most companies primarily make money by selling products (Gebauer et al., 2006). Second, the Swedish manufacturing industry significantly contributes to the country's economic growth, accounting for 20 percent of the country's GDP and providing over one million employment opportunities. The manufacturing sector has traditionally been one of the mainstays of the Swedish economy, and it continues to be a significant contributor to the country's economic growth accounting for 75% of Swedish

exports (Armelius, 2022). The Swedish manufacturing industry remains essential to the country's economy and will likely continue to drive economic growth and employment opportunities (Administration, 2022). Therefore, classifying high-growth companies in this vital sector would be beneficial for investors and policymakers to make informed decisions about the prospects of manufacturing companies in Sweden. The study fills the gap in the literature by observing big publicly traded companies from one industry. Classifying publicly traded firms would be beneficial, especially to investors, since these companies are traded on another level compared to small and medium-sized enterprises. Previous studies, such as Zhou & Gumbo (2021), have looked at small to medium-sized manufacturing enterprises and found that modern machine learning algorithms, such as SVM and ANN, outperform logistic regression regarding classification accuracy and misclassification rate. Similarly, this study employs these algorithms to classify high-growth companies, contributing to the existing literature on the binary classifying
power of classical statistical models compared to modern machine learning algorithms.

This study differs from Has et al. (2016) as the variables employed for classification are selected based on the business and economic theory of firm growth and its determinants, along with some other financial measures and margins that have shown high importance in similar studies such as Witteloostuijn & Kolkman (2019). By selecting variables based on established theories and previous research, this study aims to improve the accuracy and relevance of the classifying models.

The aim is not to investigate the causal effect of certain variables on firm growth or to test hypotheses about their marginal effects. However, some variables are included in this study due to displaying these effects in previous research. Instead, the focus is on comparing the classification accuracy as well as other measures of three models for classifying high-growth manufacturing companies on the Swedish stock market over four years

The more advanced machine learning algorithms, such as the SVM and ANN, are usually deemed to have higher accuracy in classification problems, such as classifying high-growth companies, compared to the traditional logistic regression. This hypothesis is grounded in previous literature and also partly in the theory of firm growth, which indicates that multiple determinants can have positive or negative impacts on growth depending on many factors of

the firms analysed (size, country, sector, and their interactions can be complex) (Nassar et al., 2014). Unlike logistic regression, SVM and ANN are semi-parametric models and do not impose a linear relationship between predictors and the target variable. This flexibility may allow them to capture these complex interactions more effectively (Charpentier et al., 2018). The result of this study further validates this. The SVM and the ANN has better accuracy and lower misclassification rate compared to the logistic regression on the test data, and the SVM could be argued as the best model for overall classification performance when classifying high-growth manufacturing firms in Sweden, at least when observing the results from this study.

The thesis is structured as follows: first, a section describing binary classification and prediction. Second, a literature review will be presented, comparing classical econometric models with machine learning algorithms, focusing on prediction and classification applications. Third, the theory of firm growth and its determinants will be discussed. Next, the data used in the study will be described, followed by a methodology section that outlines the three models (logistic regression, support vector machine, and artificial neural network), as well as the hyperparameter choices that have been made to build them. The results of the three models will then be presented, with interpretations provided. Finally, a discussion of the results and potential shortcomings in data, models and assumptions will conclude the thesis.

2 BINARY CLASSIFICATION AND PREDICTION

This section covers the basics of binary classification, prediction and evaluation measures of a binary classification model.

Prediction and, in turn, forecasting start with binary classification, which involves classifying instances into one of two mutually exclusive groups. In this thesis case, this would be non-high-growth firms and high-growth firms. The firms with high growth are coded as a 1, and those with non-high growth are coded as 0.

The main goal of binary classification is to create a model that can correctly classify an observation class based on its properties. There are various essential ideas and methods involved in this process.

Linear classification is a common approach in binary classification. It assumes a linear relationship between the features and the class labels. The classifier aims to find a hyperplane defined by a weight vector w and a bias term b that separates instances of different classes in the feature space. The classification rule can be expressed as equation 1 below.

$$f(x) = sign(w^T x + b) \text{ (1)}$$

Here, w^T denotes the transpose of the weight vector, and sign() assigns 1 for positive values and 0 for negative values. The decision boundary is defined as $(w^T x + b) = 0$.

In this thesis, one of the models employed is logistic regression, another prevalent technique for binary classification. Given its features, it models the probability of an instance belonging to class 1. The logistic function, or sigmoid function, is applied to the linear combination of features:

$$P(y = 1|\,\mathrm{x}) = \mathrm{sigmoid}(w^T x + b) = 1/(1 + e^{-w^T x + b}) \text{ (2)}$$

The sigmoid function maps the linear combination $(w^T x + b)$ to a value between 0 and 1, representing the probability of the instance belonging to class 1. The probability of an instance belonging to class 0 is 1- $P(y = 1|\,\mathrm{x})$. The decision boundary is usually set at a probability threshold of 0.5.

An important factor to consider is the feature selection for binary classification. Selecting or extracting features is crucial to obtain the relevant information from the data. Finding and selecting the traits that most significantly contribute to the difference between the two classes is required. Labelled data are used in the binary classification training phase. Observations containing class or category information are what constitute labelled data.

The observations in the labelled dataset have known class labels and are used to train the model. During training, the model discovers the underlying patterns and connections between the input features and the related class labels.
The model can be used to make classifications on unobserved data after training. In order to categorise the new observations into one of the two classes, the model analyses their feature

values and applies the learnt patterns. The projected class label indicates the observation's membership in the specified class.

Evaluation of binary classification models is essential to assess their performance and generalisation ability. Several evaluation metrics are commonly used (Hastie et al., 2009).
The metrics used to evaluate the three models in this thesis are accuracy, misclassification rate, specificity, sensitivity and Area under the curve score (AUC). The mathematical interpretations for the measures are in Table 1. Accuracy describes the number of correctly specified instances out of the total number of instances. Misclassification refers to the incorrect assignment of an instance to a particular class in a classification task. It occurs when a model classifies the wrong class label. Sensitivity measures the proportion of actual positive instances correctly identified as positive by the model. It quantifies the ability of the model to detect positive instances or avoid false negatives. Sensitivity is the ratio of true positives to the sum of true positives and false negatives. Specificity measures the proportion of actual negative instances correctly identified by the model. It quantifies the ability of the model to avoid false positives. Specificity is the ratio of true negatives to the sum of true negatives and false positives. AUC is the Area under the curve in a plot with the True positive rate on the y-axis and the False positive rate on the X-axis. AUC is the probability that the model ranks a random positive instance more highly than a random negative instance over all possible thresholds. An AUC of 1 is the best possible model, and an AUC of 0,5 would mean the model is as good at classifying as a random coin toss (Lindholm et al., 2022).

- True Positives (TP): The number of instances correctly classified as positive.
- False Negatives (FN): The number of instances incorrectly classified as negative.
- False Positives (FP): The number of instances incorrectly classified as positive.
- True Negatives (TN): The number of instances correctly classified as negative.

Table 1: Mathematical interpretation of the performance measures	
Performance measure	**Ratio**
Accuracy	(TP + TN) / (TP + FN + FP + TN)
Misclassification rate	(FP + FN) / (TP + FN + FP + TN)
Sensitivity (True positive rate)	TP / (TP + FN)
Specificity (1-False positive rate)	TN / (TN + FP)

3 MACHINE LEARNING VS CLASSIC ECONOMETRICS

This section explains the main distinction between classic econometric models, such as logistic regression, and modern machine learning algorithms, such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN). After that, a review of relevant literature comparing these models will be presented.

Econometric models are designed to estimate the parameters of a statistical model that best fit the data, while machine learning algorithms are first and foremost designed to learn a function that maps inputs to outputs. The main difference between classic econometric models, such as logistic regression and modern machine learning algorithms, such as SVM and ANN, is that econometric models are based on statistical inference. In contrast, machine learning algorithms are based on optimisation. In recent years, machine learning models have been found to be more effective than traditional econometric methods regarding predictions and classifications. What they gain in predictive and classification power, they, however, lose in explanatory power. They are also capable of handling much larger datasets.

Economic theory has historically served as the foundation for econometric models, which are often parametric. Thus, conventional statistical inference techniques (such as maximum likelihood and the method of moments) are applied to estimate the values of a vector of parameters. Unbiased estimators are crucial because they allow for the creation of a lower bound on the variance (the Cramer-Rao bound), just like in statistics. Taylor expansions, the Law of large numbers, and the central limit theorem are all crucial components of asymptotic theory. Contrarily, non-parametric models in machine learning are frequently constructed using

nearly entirely data (i.e., no distribution hypothesis), and the hyperparameters (cost, penalty parameter, kernel function.) are improved via cross-validation.

If a linear model can describe the connection between the variables, a well-described parametric model, such as logistic regression, should be able to make accurate classifications. In contrast, statistical methods derived from machine learning should perform better if the parametric model is incorrectly described because the connection between the variables and the class (in binary classification) is nonlinear and involves strong cross effects (Charpentier et al., 2018).

In the article "A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study" (Giovanni et al., 2020), The authors attempt to predict the financial crisis of 2008 in Italy by applying both classical econometric model such as the autoregressive model and standard ordinary least squares and compare it to several machine learning models. They found how the accuracy of the predictions is improved using the machine learning models in short-term predictions. The same conclusion has also been shown when forecasting inflation, as in "Machine learning methods for inflation forecasting in Brazil: New contenders versus Classical Models where Araujo & Gaglianone (2023) concludes that no universal best model can be chosen, but with over 50 different models tested, the Machine learning models outperform the classic econometric models in terms of mean squared error. There has also been empirical evidence for long-term forecasting superiority. In Herrera et al., 2019 the comparison is made on prices for energy commodities coming to the same conclusion.

Haselbeck et al. (2020) forecast consumer demand for horticultural products. They do this by predicting sales numbers. The results show that the nine machine learning algorithms, for example, the Lasso regression, Artificial neural net and Extreme Gradient Boosting, generally outperform the classical forecasting algorithms, such as the Exponential Smoothing and the Seasonal Autoregressive Integrated Moving Average, in terms of forecast accuracy, with a lower root mean squared error (RMSE) and mean absolute percentage error (MAPE). The article highlights the potential of machine learning models to provide more accurate and robust predictions for horticultural sales, which can help growers and suppliers to optimise their production and logistics planning.

Using demographic and financial performance Witteloostuijn & Kolkman (2019) compared a random forest analysis to a regular regression to compare the goodness of fit on firms originating from Belgium and the Netherlands. Their data had high variations in the growth rates, making the prediction challenging to capture with linear regression. They found that the random forest, a standard machine learning algorithm, fared three to four times better than the regular regression. The authors also concluded that their findings indicated that firm growth might not be as random as previously thought.

4 FIRM GROWTH AND IT IS DETERMINANTS

This section covers the literature and theory on firm growth and the possible determinants of firm growth and is the foundation for some of the variables used in the comparisons.

4.1 Size

Gibrat's Law implies that firm growth is a random process, connoting that firm growth rate is similar for all enterprises in the market (Geroski, 2005; Stam, 2010). Put differently, Gibrat's Law implies that a firm's growth process is stochastic and not determined by internal or external drivers. However, subsequent studies have tested and largely rejected the validity of this theoretical model, as in Nassar, Almsafir, & Al-Mahrouq, (2014). It was found that in most of the manufacturing sector, Gibrat's Law fails to hold, but for the service sector, Gibrat's Law was valid. Furthermore, most empirical studies applied in developed countries rejected Gibrat's Law. In the manufacturing sector in a developed country like Sweden, we would assume that the growth should be influenced by the size of the enterprise based on this information and should be a valid variable to classify high-growth firms. Building a model that can classify and predict firm growth should be challenging if it is a random process.

4.2 Age

The impact of firm age on growth rates is a topic of ongoing debate. Choi (2010) and Evans (1987) found evidence supporting the notion that younger firms have higher growth rates than older firms. This finding is consistent with the results obtained by Fizaine (1968), who conducted one of the earliest investigations into the relationship between firm age and growth.

Fizaine analysed the growth of establishments in the French region of Bouches-du-Rhone and observed a negative effect of age on growth. Furthermore, Fizaine found that the variance of growth rates tends to decrease with age.

Numerous studies have also demonstrated a negative effect of age on growth at the firm level. For example, Variyam and Kraybill (1992) investigated US manufacturing firms and found that age negatively correlates with growth. Similarly, Geroski and Gugler (2004) examined large European companies, and Yasuda (2005) studied Japanese manufacturing firms, which also found that age negatively affects growth. These findings suggest that younger firms have higher growth rates than older firms across different countries and industries.

4.3 Macroeconomic Factors

The empirical literature on firm growth rates suggests that variation in firm growth is more significant across industries than across countries. Nonetheless, investigating the influence of macroeconomic factors on firm growth rates provides valuable insights. Researchers have examined the relationship between firm growth and the business cycle. Higson et al. (2002, 2004) analyses data on US and UK firms spanning over three decades and find that mean growth rates are sensitive to macroeconomic fluctuations.

Cross-country differences in firm growth rates have also been explored. Beck et al. (2005) conducted a study analysing a size-stratified firm-level survey database encompassing over 4,000 firms in 54 countries. They observe that firms in bigger, wealthier, and faster-growing countries tend to experience significantly higher growth rates. Additionally, they find a positive correlation between GDP growth rate and firm growth, indicating that firms tend to grow faster in economies with more significant growth opportunities.

The impact of inflation on growth rates was also investigated in the study by Beck et al. (2005). The authors cautioned that their findings regarding a positive relationship between inflation and growth rates might reflect that firm sales growth is typically reported in nominal terms.

4.4 Financial Performance

Coad (2005) finds a statistically significant relationship between profit rate and sales growth for French manufacturing firms using OLS but no effect for the non-parametric analysis. Nevertheless, the magnitude of the coefficient is so tiny that he concludes that a firm's profit

rate and growth rate are entirely independent. Bottazzi et al. (2006) find similar results in their analysis of Italian firms. However, in some empirical literature, these metrics have been shown to impact the prediction capabilities of the model used (Witteloostuijn & Kolkman, 2019).

5 METHODOLOGY

This section presents a theoretical foundation for the three models employed in this study, including an explanation of their fundamental principles. Additionally, it discusses the selection of hyperparameters associated with each model.

All models are built and analysed in R, the packages used include: plm,glm,
Caret, e1071, neuralnet.

5.1 Logistic Regression

Logistic regression is a binary classification model that predicts the class of instances using maximum likelihood estimation (Lussier, 1995). This model has been widely used in modelling firm growth, as Megaravalli (2017) demonstrated.

$$\log\left[\frac{P(X)}{1-P(X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (3)$$

In the logistic regression model, p(X) represents the probability of the outcome (high-growth), β0 is the intercept, and β1, β2,…, βp are the model coefficients. X1,X2,…,Xp represent the drivers of growth performance (Equation 3). When applied to training and test datasets, the trained logistic regression model produces results ranging between 0 and 1, where 0 indicates non-high-growth, and 1 indicates high-growth. A cutoff point of 0.5 is used to determine whether a manufacturing firm in Sweden experienced a high-growth rate between 2018-2021. This means that a manufacturing firm with a p(X) value equal to or greater than 0.5 is classified as a high-growth firm, while those with a value below 0.5 are classified as non-high-growth firms.

5.2 Support Vector Machines

Support Vector Machine is a popular machine learning algorithm commonly used in binary classification problems. It excels in handling high-dimensionalities in the data and works very well with non-linearly separable data (Cortes & Vapnik, 1995; Vapnik, 1998). It aims to find a hyperplane that best separates the data points into two classes. The algorithm maximises the margin between the hyperplane and the closest data points of each class, known as the support vectors (Cortes & Vapnik, 1995).

Selecting a kernel function is one of the essential choices while creating a Support Vector Machines model. According to Boser et al. (1992), the kernel function transforms the input space's data points into a high-dimensional feature space where the data points are more likely to be linearly separable. Various kernel functions are available, including sigmoid, radial basis function (RBF), linear, and polynomial. The dataset and the issue being solved will determine which kernel function is used. The general kernel SVM is demonstrated by Awad and Khanna (2015) as per Equation 3:

$$\mathrm{K}(x,u) = \sum_r \varphi_r(x)\, \varphi_r(u) \quad (4)$$

$\mathrm{K}(x,u)$: This represents the kernel function, which takes two input instances, x and u, which are vectors and computes a similarity measure between them. $\varphi_r(x)$ and $\varphi_r(u)$ term represents the mapping of instance x respectively u into the r-th dimension or feature in the feature space. r typically corresponds to the number of dimensions or features in the feature space.

When the data can be separated linearly, the linear kernel function is frequently utilised since it generates the optimal hyperplane for separating the data points. Data that cannot be separated linearly can be handled using the polynomial kernel function by projecting the data points into a higher-dimensional space.

The RBF kernel function is frequently used for non-linearly separable data because it supports intricate decision boundaries. Another alternative for non-linearly separable data is the sigmoid kernel function; however, it is less utilised than the other kernel functions (Boser et al., 1992; Cortes & Vapnik, 1995).

The cost parameter C (see Equation 6) determines the trade-off between maximising the margin and reducing the classification error. If C is greater, the margin will be less, but there will be fewer misclassifications, while if C is lower, the margin will be bigger, but there will be more misclassifications.

The choice of kernel function and cost parameter will depend on the specific dataset and the problem being solved, and it may require some experimentation to find the optimal values. For this analysis, both the polynomial and the linear function were tested in the model. The linear function performed better and is used for the model comparison. Using cross-validation, the cost parameter chosen is 1.

$$\mathrm{K}(x, u) = x^T \cdot u \quad (5)$$

$$C * [\sum(\max(0, 1 - y_i * (w^T x_i + b))) + \lambda * ||\mathrm{w}||^2] \quad (6)$$

The linear kernel (Equation 5) computes the dot product between the feature vectors x and u. Geometrically, the linear kernel measures the similarity or the degree of alignment between the two instances. It captures the notion of similarity based on the linear relationship between the features of the instances. Example: the vectors x and u represent the predictors for two manufacturing firms, Firm 1 and Firm 2. The dot product of these vectors will be large if the firms exhibit high similarity in their predictors, indicating common characteristics.
Conversely, the dot product will be small if the firms are dissimilar, suggesting differences in their predictor patterns.

The cost function (6) guides the training process to find the optimal hyperplane in the transformed feature space. C is the regularisation parameter, determining the trade-off between maximising the margin and minimising the classification errors.
Σ represents the sum over all training instances.
y_i is the class label of the i-th training instance(0 or 1).
w and b are the hyperplane's weight vector and bias term, respectively.
x_i is the feature vector of the i-th training instance.
λ is the regularisation parameter that controls the penalty for the magnitude of the weight vector w. It helps prevent overfitting by discouraging large weights.

Support Vector Machines (SVM) is an effective categorisation method for assessing business expansion (Zhou & Gumbo, 2021). It is desirable for predictive modelling jobs since it can handle non-linearly separable data and is resilient to noise. Based on this information, it should attain high levels of classification accuracy for firm growth (Awad & Khanna, 2015).

5.3 Artificial Neural Networks

Artificial neural networks (ANNs) are a machine learning technology that has gained popularity for resolving challenging classification issues like binary classification. ANNs, which can simulate nonlinear interactions between input and output variables, were inspired by the human brain and nervous system. This section will review some of the many decisions that must be made while creating an ANN for binary classification. The most important are network architecture, activation function and training algorithm (Goodfellow et al., 2016); LeCun et al.,2015); Shalev-Shwartz & Ben-David, 2014).

Network architecture is the quantity and configuration of nodes or neurons within an ANN. Each layer of nodes in an ANN's architecture can comprise one or more neurons. While the output layer generates the predictions for the network, the input layer receives the input variables. Intermediary layers, called hidden layers, take on the characteristics of the input variables. Given that x_i is the ith input to the ANN node, w_i the ith input weight, n the number of inputs, b the bias term and o the node output. Then to resolve a classification problem, Equation 7 is used and represents the Equation for one node in the network.

$$o = \sigma \sum_{i=1}^{n} (wixi + b) (7)$$

$$\text{Where; } \sigma(x) = \frac{1}{1+e^{-x}} (8)$$

In this Equation, the sigmoid function (Equation 8) takes the weighted sum of inputs (represented as $\sum_{i=1}^{n} (wixi + b)$) as its argument and outputs a value between 0 and 1. The weighted sum is calculated by summing up the products of the input values (xi) and their corresponding weights (wi) and adding the bias term (b).

Equation (5) represents the basic computation performed by a single neuron in an ANN. In a neural network with multiple layers, the outputs of the neurons in one layer serve as inputs to the neurons in the next layer, and this process continues until reaching the output layer, where the final predictions or outputs of the network are obtained.

A straightforward architecture with one or two hidden layers is frequently adequate for binary classification. Starting with a minimal number of hidden layers and gradually increasing them until the model performance stabilises or declines is advised by Goodfellow et al. (2016). The risk of overfitting the training data, which can result in subpar generalisation performance on new data, also rises with the number of layers.

The network gains nonlinearity via activation functions, which enables it to describe intricate interactions between input and output variables. The data's nature and the task's difficulty influence the choice of the activation function. The ReLU (rectified linear unit) and sigmoid activation functions are frequently used.

The sigmoid activation function is frequently employed in the output layer for binary classification since it generates a probability between 0 and 1, which may be thresholded to generate a binary prediction, ReLU or its variants are frequently chosen for the hidden layers since they are computationally effective and can avoid the vanishing gradient problem that might happen when using sigmoid or tanh. ReLU has been demonstrated to function effectively in reality and is a good default solution for many issues (LeCun et al., 2015).

Based on the training data, optimisation algorithms are utilised to adjust the weights and biases of the network. ANNs frequently employ gradient-based methods like backpropagation. Concerning the network parameters, these methods compute the gradient of the loss function and update the network parameters accordingly.
The binary cross-entropy loss function is a popular choice for binary classification since it is suitable for probabilistic predictions. The size of the dataset and the model's complexity affect the optimisation algorithm that is selected. While Adam and RMSprop are chosen for larger datasets or more sophisticated models, stochastic gradient descent (SGD) is a common

approach for small to medium-sized datasets. A thorough description of optimisation techniques for machine learning is given by Shalev-Shwartz & Ben-David (2014).

Several choices must be made regarding network architecture, activation functions, and training algorithms to build a practical ANN for binary classification. A simple architecture with two hidden layers, the sigmoid activation function in the output layer, ReLU in the hidden layers, and a gradient-based optimisation algorithm, such as SGD, can produce good results for many binary classification problems. Hence, these choices have been made in this analysis. There is also the concept of nodes in the layers of the neural network. The “Rule of thumb” (Bishop, 2011) was applied for this purpose, using two nodes per layer.

6 DATA

This section offers an overview of the data utilised in this thesis, including its collection process, pre-processing steps, and partitioning methodology.

The data set consists of yearly financial and descriptive information for 163 publicly traded manufacturing companies in Sweden spanning nine years, from 2014 to 2022. The 163 companies were isolated from the rest of the Swedish stock market by NACE code. NACE (Nomenclature statistique des activités économiques dans la Communauté européenne) is a statistical classification system used to categorise economic activities in the European Union (EU) and other countries that use the system. NACE classifies the companies that are in the dataset as manufacturing companies.

Table 2: Variable Description	
Variable (Predictors)	**Description**
Revenue	Total income generated by the manufacturing firm during a year.
Number of employees	Total number of individuals employed by the company that year.
Age of Company	Years that have elapsed since the company was founded (current year-year of foundation).
Working Capital	Difference between current assets and current liabilities
Total Assets	Total value of the company's assets, tangible and intangible assets.
Total Equity	The total value of the company's assets subtracting its liabilities.
Current Ratio	The ratio of the company's current assets to current liabilities.
GDP growth rate	Percentage change in the value of goods and services produced by the Swedish economy.
Inflation	The rate at which the general level of prices for goods and services is increasing, resulting in a decrease in the purchasing power of a currency.
Interest rate	The percentage charged by a lender to a borrower for the use of money (Styrräntan).
Variable (Dependent)	**Description**
High-growth-Yes-No (HGYN)	Firms with revenue growth of 20% or higher between 2018-2021 are coded as 1, while those below the 20% threshold are coded as 0.

The choice of Revenue, Number of employees, and Age of the Company is due to these three variables' presence in economic theory as described above. Working Capital Total Assets, Total Equity and Current Ratio are included due to their prediction importance in the study by Witteloostuijn & Kolkman (2019). GDP growth rate, Inflation and Interest rate were also included to cover the potential influence of macroeconomic factors on the firm's growth.

The dataset was obtained from Bloomberg. The unedited panel was unbalanced, containing many missing observations for specific years and companies. Any missing or incomplete data was removed to ensure panel data completeness and balance. The resulting panel dataset used includes four years from 2018 to 2021. Both the time dimension and cross-sectional dimension of the panel align with previous studies (Zhou & Gumbo, 2021; Almsafir et al., 2015; Hermelo & Vassolo, 2007). A total of 10 predictors and one binary target variable are included in the data set (see Table 2). The sample size of the manufacturing firms is a total of 156 unique firms after removing missing values.

6.1 Dimension Reduction

Dimension reduction is performed via principal component analysis (PCA) to reduce the multicollinearity between the variables (see Table 5 in the appendix).

PCA is a dimensionality reduction technique widely used in data analysis and machine learning. Its primary objective is to transform a high-dimensional dataset into a lower-dimensional representation while retaining as much information as possible. PCA achieves this by identifying a new set of variables, called principal components, that are linear combinations of the original variables.

The key idea behind PCA is to capture the maximum variance in the data with the first few principal components. The first principal component explains the largest variance, followed by the second component, and so on. Each principal component is orthogonal to the others, meaning they are uncorrelated. By selecting a subset of principal components, it is possible to reduce the dimensionality of the dataset while preserving most of its information.
PCA is beneficial for several reasons. Firstly, it simplifies the analysis by reducing the number of variables, making it easier to visualise and interpret the data. Secondly, it can uncover the underlying structure or patterns in the data, revealing relationships between variables that may not be apparent initially. Additionally, PCA can remove redundant or irrelevant features, improving computational efficiency and reducing the risk of overfitting in subsequent analyses or models. One potential limitation of PCA is the decreased interpretability of individual variables meaning that it becomes challenging to draw specific inferences based on the principal components themselves (Jolliffe, 2002). However, interpretability is not an objective

in the context of this thesis. The main focus is on evaluating the classification performance of the models.

Before PCA is performed, the variables are standardised to have a mean of zero and a variance of 1, as is the norm, due to the technique not being scale invariant (Jolliffe, 2002).
Three principal components are chosen using the most common rule, the Kaiser rule (Kaiser, 1960). These components are used for classification and performance assessment using Logistic regression, SVM, and ANN algorithms.

6.2 Data Partitioning

In machine learning, it is crucial to partition the dataset into two parts: the training (in-sample) and test (out-of-sample) data sets. The training data fits the model, while the test data is used for model validation or testing. This procedure ensures that the model performance on unseen labelled data is examined, preventing overfitting to the training data. For this study, a 70:30 training: testing data split ratio is used, similar to a related study by Delen et al. (2013), to enable a more accurate comparison with the results of Zhou & Gumbo (2021), who also used this split. The three machine learning algorithms are fitted using the training data and then tested on the test set. The performance of each algorithm is evaluated based on their classification performance on the test data.

7 RESULTS

This section outlines the obtained results from comparing the three models using the abovementioned data. It begins by presenting the Confusion Matrix and subsequently discusses various performance measures.

The confusion matrix presented in Table 1 illustrates the performance of three different classification algorithms, namely Logistic Regression, Support Vector Machine (SVM), and Artificial Neural Network (ANN), on both the train and test datasets. This matrix provides valuable insights into the effectiveness of these models in predicting the outcome classes.

Table 3: Confusion Matrix, Train and Test Dataset									
Output based on training data set						Output based on test data set			
		Not-High Growth(0)	High Growth(1)			Not-High Growth(0)	High Growth(1)		
Algorithm									
Logistic	Not-High Growth(0)	221	47	Correct	359	80	23	Correct	144
	High Growth(1)	31	138	Wrong	78	20	64	Wrong	43
SVM	Not-High Growth(0)	231	48	Correct	368	90	23	Correct	154
	High Growth(1)	21	137	Wrong	69	10	64	Wrong	33
ANN	Not-High Growth(0)	223	38	Correct	370	83	20	Correct	150
	High Growth(1)	29	147	Wrong	67	17	67	Wrong	37

Observing the test data set which is the main focus. The Logistic Regression model correctly classifies 80 instances as Not-High-Growth (true negatives) and 20 instances as High-Growth (true positives) in the test dataset. However, it also misclassifies 23 instances as High-Growth when they were Not-High-Growth (false positives) and 64 instances as Not-High Growth when they were High-Growth (false negatives).

The SVM model accurately classifies 90 instances as Not-High-Growth and 10 instances as High-Growth. Nevertheless, it had 23 false positive classifications and 64 false negative classifications.

The ANN correctly classifies 83 instances as Not-High-Growth and 17 instances as High-Growth. However, it did have 20 false positive classifications and 67 false negative classifications. Overall, all three models can make more correct class predictions than incorrect on both the train and test data, which is desirable. It is, however, difficult to compare the models based only on these numbers.

Table 4 below shows the performance of three different models (logistic regression, SVM, and ANN) on the training and test datasets.

Table 4: Models Evaluation on Training and Testing Dataset						
	Evaluation based on train Data			Evaluation based on test Data		
	Logistic	SVM	ANN	Logistic	SVM	ANN
Measure						
Accuracy	0,8215	0,8421	0,8467	0,7701	0,8235	0,8021
Misclassification	0,36	0,16	0,1533	0,2299	0,1765	0,1979
Specificity	0,7459	0,7405	0,7946	0,7356	0,7356	0,7701
Sensitivity	0,877	0,9167	0,8849	0,8	0,9	0,83
AUC				0,8285	0,8597	0,8502

When considering the test data, the Logistic Regression model achieved an accuracy of 0.7701, indicating that it correctly classifies 77.01% of the instances. The SVM model performed slightly better, with 82.35% correctly classified instances. The ANN model correctly classified 80.21% of the instances.

The misclassification rate, representing the proportion of misclassified instances, was 0.2299 for the Logistic Regression model, indicating that it misclassifies 22.99% of the instances. The SVM model had a lower misclassification rate of 17.65%. The ANN model performed slightly worse than the SVM, with a misclassification rate of 19.79%.

In terms of specificity, which measures the ability to identify the negative class correctly, all models perform reasonably well. The specificities of the Logistic Regression, SVM, and ANN models are comparable, indicating their ability to correctly identify instances that do not belong to the high-growth category.

When considering sensitivity, which measures the ability to identify the positive class correctly, the SVM model achieves the highest sensitivity, correctly identifying 90% of the high-growth instances. The ANN model has the highest specificity for test data (77.01%), implying it is the best model for classifying the negative class (non-high-Growth).

The SVM model has the highest AUC score (0,8597) among the three models, followed closely by the ANN model (0,8502). The AUC score suggests that the SVM model can better distinguish between positive and negative classes than the logistic regression and ANN models, regardless of the threshold. It is important to note that the difference in the AUC score between the models is relatively small. However, the results indicate that these models are arguably reliable at classifying high-growth firms from not-high-growth ones.

The differences between the train and test performances of the models are not too significantly different to assume that the models highly overfit the data. However, this is not conclusive evidence alone for that to be the case.

The conclusion could be that the best-performing model is the SVM having the best measures on both the train and test data set in all categories except for specificity, where the logistic regression slightly performs better.

The results of these measures align with the findings of Zhou & Gumbo (2021), Who also found the SVM model to perform the best on the test data, followed by the ANN and the logistic regression.

8 DISCUSSION

This section discusses the result and their implications for growth theory and existing literature. Additionally, it includes a segment highlighting the practical applicability of the models for investors and policymakers.

The findings from the model comparison align with previous literature. The two modern machine learning algorithms, SVM and ANN, outperform the logistic regression in classifying high-growth and non-high-growth companies when applied to new data. The SVM and ANN also exhibit lower misclassification rates, indicating fewer classification errors. However, it is essential to acknowledge that this classification power comes with inevitable trade-offs. Specifically, when employing dimension reduction techniques such as PCA and utilising semi-

parametric models like ANN and SVM, it becomes challenging to measure the individual variable effects on the dependent variable (Jolliffe, 2002; Ribeiro et al., 2016).
Consequently, we cannot determine the marginal effect of any of the original ten predictors on whether a manufacturing firm will exhibit high-growth. This limitation stems from the "black box" nature of modern machine learning algorithms, which lack transparency compared to logistic regression. As a result, the trustworthiness of the information derived from these predictions is diminished, reducing the value of the models for investors and policymakers (Ribeiro et al., 2016). We cannot draw firm conclusions regarding specific firm growth theories, such as Gibrat's Law, due to the inability to determine the individual effects of variables like Revenue or Number of Employees on firm classification. However, the results indicate a higher than 50/50 likelihood of correctly classifying a company's growth status for all models. This suggests that the random process theory of firm growth (Geroski, 2000) does not hold, at least for Swedish publicly traded manufacturing firms, from 2018 to 2021. In overall performance, the SVM model was the best for classifying high-growth manufacturing firms.

9 SHORTCOMINGS

This section discusses the limitations of the thesis and the reasons behind their non-addressal, and potential strategies that could have been employed in an ideal setting to mitigate these limitations.

9.1 NANCE Code Drawbacks

Using NANCE to identify manufacturing companies has its drawbacks. The ideal approach for selecting companies for the panel would involve examining all available companies individually and establishing a threshold for the proportion of revenue allowed from sources other than product sales. However, due to time constraints and the limitations of the available data, this was not feasible and carried its own limitations. Some of the companies included in the panel were initially manufacturing companies but have since transitioned into a more ambiguous area. For example, Ericsson AB is now considered a telecommunications company rather than a pure manufacturing company. Another issue with using NANCE is the wide variation in the types of products being manufactured. Manufacturing pharmaceutical

equipment, for instance, may differ significantly from manufacturing cars (such as Volvo and AstraZeneca), yet both fall under the same NANCE code.

This variation in manufacturing sectors can lead to differences in resource requirements and might skew the results. However, it is essential to note that these limitations in data grouping do not apply to most companies in the panel and are likely to have a marginal impact on the study results. Nonetheless, it suggests the findings may be more generalisable to other industries and sectors. However, this conclusion is not definitive, and further testing on other datasets would be necessary to validate it. An improvement could be achieved by analysing each company individually and considering various types of manufacturing, such as on-demand manufacturing or pharmaceuticals, and controlling for this variable in the models, similar to the approach used by Witteloostuijn and Kolkman (2019), who categorised firms based on the second-level NACE code.

9.2 Data Limitations

Overall, several limitations are associated with the data used in this study. Firstly, the sample of firms is relatively small as it only includes companies listed on the Swedish stock market. Although this is consistent with other studies (Zhou & Gumbo, 2021), it restricts the generalizability of the findings. Machine learning algorithms like SVM and ANN are designed to handle large datasets with thousands or even millions of observations, so the limited sample size is a drawback in this study.

Furthermore, the sample of companies is skewed towards larger, more mature publicly traded firms. It would be beneficial to include information about smaller and medium-sized manufacturing firms in Sweden to provide more comprehensive models and better compare the results with previous studies like Zhou & Gumbo (2021).

Another limitation is the relatively short time frame of four years. While this duration aligns with other studies (Zhou & Gumbo, 2021; Has et al., 2016), it restricts the ability to capture long-term trends and cyclical patterns. Panel data analysis can offer advantages such as generalizability and capturing temporal dynamics, but a longer period would enhance these aspects (Lu & Su, 2020). Ideally, the models should be tested on several observations spanning several years to obtain more robust results.

In summary, the limitations of the data, including the small sample size, limited selection of firms, and short time frame, should be considered when interpreting this study's results.

It is important to acknowledge that the data used in this study may contain various biases that could influence the results, with outliers being one potential concern. However, due to the already limited amount of data in the panel, the decision was made not to remove outliers. It should be noted that both the ANN and SVM models are generally considered less effective at handling outliers (Hastie et al., 2009), which could potentially impact the results negatively.

Although PCA was employed to mitigate multicollinearity, it is not a definitive solution to the problem. If multicollinearity exists, the models may overfit the data, reducing generalizability and predictive performance (Jolliffe, 2002). Additionally, other biases, such as violations of the independence of errors, heteroscedasticity, and autocorrelation, which primarily affect the inferential aspects of the models, can also have implications for their classification ability(Verbeek, 2008).

It is worth mentioning again that while this thesis focuses primarily on classification rather than inference, these biases can still affect the classification ability of the models. These biases can result in more misclassification and less accuracy, thereby diminishing the overall performance and trustworthiness of the models in making future classifications (Verbeek, 2008). A comprehensive robustness check was not conducted, partly due to the primary negative impact of these biases on the inference of the model and partly due to time constraints and their absence in similar studies (Zhou & Gumbo, 2021; Has et al., 2016; Witteloostuijn & Kolkman, 2019).

9.3 Model Tuning Limitations

There are also limitations in the tuning of the models. While the methodology section explains the hyperparameters considered in the models, it is essential to note that these complex models have numerous other hyperparameters that can impact their performance. To achieve the best possible results, extensive time and effort would be required to thoroughly cross-validate and fine-tune each component of the models (Bishop, 2011).
Due to time constraints and the scope of the research, a comprehensive tuning of all model hyperparameters was not conducted in this study. Therefore, it is possible that the results could

have been further improved if more time and resources were available for thorough hyperparameter optimisation.

10 CONCLUSION

In conclusion, the model comparison findings are consistent with existing literature, highlighting the superior classification capabilities of SVM and ANN over logistic regression in classifying high-growth and non-high-growth companies. However, it is essential to acknowledge the limitations associated with these models. Dimension reduction techniques like PCA and adopting semi-parametric models introduce challenges in quantifying the individual effects of predictors on the dependent variable, hampering our understanding of their marginal impact. The "black box" nature of modern machine learning algorithms makes them untrustworthy, diminishing their value for investors and policymakers. While firm growth theories such as Gibrat's Law cannot be conclusively validated due to the inability to determine variable effects, the study indicates a non-random nature in classifying firm growth, challenging the notion of random process theory at least for Swedish publicly traded manufacturing firms from 2018 to 2021. Future research should explore these findings on more types of firms, time spans and countries.

11 BIBLIOGRAPHY

Administration, I. T. (den 25 August 2022). *Sweden - Country Commercial Guide* . Hämtat från International Trade Administration : https://www.trade.gov/country-commercial-guides/sweden-advanced-manufacturing

Almsafir, M., Nassar, I., Al-Mahrouq, M., & Hayajneh, J. (2015). The Validity of Gibrat's Law: Evidence from the Service Sector in Jordan. *lirocedia Economics Finance*, 1602-1606.

Araujo, G., & Gaglianone, W. (2023). Machine learning methods for inflation forecasting in Brazil: New contenders versus Classical Models. *Latin American Journal of Central Banking, 4*(2).

Armelius, H. (den 2 March 2022). *Industriproduktionens sammansättning*. Hämtat från Ekonomifakta: https://www.ekonomifakta.se/fakta/ekonomi/produktion-och-investeringar/industriproduktionens-sammansattning/

Azaro, K., Djajanto, L., & Sari, P. (2019). The Influence of Financial Ratios and Firm Size on. *Advances in Economics, Business and Management Research, 136*.

Beck, T., Demirguc-Kunt, A., & Maksimovic, V. (2005). Financial and Legal Constraints to Growth: Does Firm Size Matter? *Journal of Finance, 60*(1), 137-177.

Bishop, C. (2011). *Pattern Recognition and Machine Learing.* SPRINGER-VERLAG NEW YORK INC.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifier. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Bottazzi, G., Secchi, A., & Tamagni, F. (2006). Productivity, Profitability and Financial Fragility: Evidence from Italian Business Firms. *DRUID Summer Conference.* Copenhagen.

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(7), 1-27.

Charpentier, A., Flachaire, E., & Ly, A. (2018). Econometrics and Machine Learning. *ECONOMIE ET STATISTIQUE / ECONOMICS AND STATISTICS*, 505-506.

Choi, B. (2010). The U.S. Property and Liability Insurance Industry: Firm Growth, Size, and Age. *Risk Management and Insurance Review, 13*, 207-224.

Christianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge: Cambridge University Press.

Coad, A. (2007). Testing the principle of 'growth of the fitter': the relationship between profits and firm growth. *Structural Change and Economic Dynamics, 18*(3), 370–386.

Delmar, F., Davidsson, P., & Gartner, W. (2003). Arriving at the High-Growth Firm. *Journal of Business Venturing, 18*, 189-216.

Evans, D. (1987). Tests of Alternative Theories of Firm Growth. *Journal of Political Economy, 95*(4), 657-892.

Evans, D. (1987). The Relationship between Firm Growth, Size and Age: Estimates for 100 Manufacturing Industries. *Journal of Industrial Economics, 35*, 567-581.

Fizaine, F. (1968). ANALYSE STATISTIQUE DE LA CROISSANCE DES ENTREPRISES SELON L'AGE ET LA TAILLE. *Revue d'économie politique, 78*(4), 606-620.

Gebauer, H., Friedli, T., & Fleisch, E. (2006). Success factors for achieving high service revenues in manufacturing companies. *Benchmarking An International Journal*.

Geroski, P. (2000). *The growth of firms in theory and in practice.* CEPR Press Discussion Paper No. 2092.

Geroski, P. (2005). Understanding the Implications of Empirical Work on Corporate Growth Rates. *Managerial, 26*(2), 129-138.

Geroski, P., & Gugler, K. (2004). Corporate growth convergence in Europe. *Oxford Economic Papers, 56*(4), 597-620.

Giovanni, C., Inserra, G., & Limosani , M. (2020). A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study. *Mathematics, 2*(8).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning series).* MIT Press.

Has, A., Zekic-Susac, M., Sarlija, , N., & Bilandzic, A. (2016). Predicting company growth using logistic regression and neural networks. *Croatian Operational Research Review, 7*, 229/248.

Haselbeck, F., Killinger, J., Menrad, K., Hannus, T., & Grimm, D. (2020). Machine learning outperforms classical forecasting on horticultural sales predictions. *Machine Learning with Applications, 7*.

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

Henrekson, M., & Johansson, D. (2010). Gazelles as job creators: a survey and interpretation of the evidence. *35*(2), Small Business Economics.

Hermelo, D., & Vassolo, R. (2007). The determinants of firm's growth: an emliirical examination. *Revista Abante*, 3-20.

Herrera, G., Constantino, M., Tabak, B., Pistori, H., Su, J.-J., & Naranpanawa, A. (2019). Long-term forecast of energy commodities price using machine learning. *Energy, 179*, 214-221.

Higson, C., Holly, S., & Kattuman, P. (2002). The cross-sectional dynamics of the US business cycle: 1950-1999. *Journal of Economic Dynamics and Control, 26*, 1539-1555.

Higson, C., Holly, S., Kattuman, P., & Platis, S. (2004). The Business Cycle, Macroeconomic Shocks and the Cross-Section: The Growth of UK Quoted Companies. *Economica, 71*, 299-318.

Hinton, G., Bengio, Y., & LeCun, Y. (2015). Deep Learning. *Nature*, 436-444.

Jolliffe, I. T. (2002). *Principal Component Analysis.* New York: Springer Science & Business Media.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 203-2015.

Khanna , R., & Awad, M. (2015). *Support Vector Machines for Classification: theories, Concepts, and Applications for Engineers and System Designers.* Apress.

Kraybill, D., & Variyam, J. (1992). Empirical evidence on determinants of firm growth. *Economics Letters, 38*(4), 31-36.

Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. B. (2022). *Machine Learning - A First Course for Engineers and Scientists.* Cambridge: Cambridge University Press .

Lu , X., & Su, L. (2020). Determining individual or time effects in panel data models. *Journal of Econometrics*, 60-83.

Lussier, R. (1995). A Nonfinancial Business Success versus Failure Prediction Model for Young Firms. *Journal of Small Business Management, 33*(1), 8-20.

Mason, C., & Brown, R. (2014). ENTREPRENEURIAL ECOSYSTEMS AND GROWTH ORIENTED ENTREPRENEURSHIP Background paper prepared for the workshop organised by the OECD LEED Programme and the Dutch Ministry of Economic Affairs on.

ENTREPRENEURIAL ECOSYSTEMS AND GROWTH ORIENTED ENTREPRENEURSHIP (ss. 1-39). The Hague: OECD.

Megaravalli, A. (2017). Estimating growth of SMES using a logit model, evidence from manufacturing companies in Italy. *Management Science Letters,, 7*(3), 125-134.

Nassar , I., Almsafir, M., & Al-Mahrouq, M. (2014). The Validity of Gibrat's Law in Developed and. *Procedia-Social Behavioral Sciences, 129*, 266-273.

OECD. (2021). *Measuring distance to the SDG targets 2021: An assessment of where OECD countries stand.* OECD.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ss. 1135–1144). New York: Association for Computing Machinery.

Shalev-Shwartz, S., & Shai, B.-D. (2014). *Understanding Machine Learning.* Cambridge: Cambridge University Press.

Stam, E. (2010). Growth beyond Gibrat: firm growth processes. *Small Bus Economics, 35*, 129–135.

Verbeek, M. (2008). *A Guide to Modern Econometrics.* West Sussex: Wiley.

Virtanen, J. (2019). PREDICTING HIGH-GROWTH FIRMS WITH MACHINE LEARNING METHODS. *Unpublished master's thesis*.

Witteloostuijn, A., & Kolkman, D. (2019). Is firm growth random? A machine learning perspective. *Journal of Business Venturing Insights, 11*.

Yasuda, T. (2005). Firm Growth, Size, Age and Behavior in Japanese Manufacturing. *Small Business Economics, 24*, 1-15.

Youn, H., & Gu, Z. (2010). Predicting Korean lodging firm failures: An artificial neural network model along with a logistic regression model. *International Journal of Hospitality Management, 20*(1), 120-127.

Zhou, H., & Gumbo, V. (2021). Comparative Analysis of A Traditional and Machine Learning Techniques in Predicting SMMES Growth Performance. *Academy of Entrepreneurship Journal, 27*(3).

12 APPENDIX

12.1 Correlation Matrix

Table 5: Correlation matrix

	HGYN	Revenue	Numberofemployees	AgeofCompany	WorkingCapital	TotalAssets	TotalEquity	Currentratio	GDP Growth	Inflation	InterestRate
HGYN	1										
Revenue	-0,2401*	1									
Numberofemployees	-0,2982*	0,7279*	1								
AgeofCompany	-0,2031*	0,3916*	0,3738*	1							
WorkingCapital	-0,2061*	0,8242*	0,6306*	0,4521*	1						
TotalAssets	-0,2349*	0,9772*	0,7357*	0,4169*	0,8491*	1					
TotalEquity	-0,2202*	0,897*	0,675*	0,5062*	0,8318*	0,9481*	1				
Currentratio	0,0463	-0,1068*	-0,1289*	-0,1869*	-0,0677*	-0,1096*	-0,1227*	1			
GDP Growth	0	-0,0077	0,0017	-0,0294	-0,0285*	-0,0305	-0,0547	0,0317	1		
Inflation	0	-0,0032	0,0038	-0,0286	-0,0313	-0,0265	-0,0479	0,0261	0,9342*	1	
InterestRate	0	-0,0036	0,0034	-0,0195	-0,0272	-0,019	-0,0309	0,0266	0,6178*	0,7793*	1

* indicates significance at 5% level

12.2 Model Evaluation without dimension reduction

Table 6: Models Evaluation on training and testing data, all ten predictors (no dimension reduction)						
	Evaluation based on train data			Evaluation based on test data		
	Logistic	SVM	ANN	Logistic	SVM	ANN
Measure						
Accuracy	0,625	0,634	0,508	0,572	0,599	0,492
Misclassification	0,375	0,366	0,492	0,428	0,401	0,508
Specificity	0,692	0,930	0,508	0,655	0,920	0,483
Sensitivity	0,575	0,417	0,508	0,500	0,320	0,500