Predicting movements in economic growth based on Indicators of Educational Development using historical data

Maud Calame
ANR: U653034

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COMMUNICATION AND INFORMATION SCIENCES,
MASTER TRACK DATA SCIENCE BUSINESS & GOVERNANCE,
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

Thesis committee:

Dr. Ç. Güven
Dr. T.J. Wiltshire

Tilburg University
School of Humanities and Digital Sciences Department of Cognitive Science & Artificial Intelligence Tilburg, The Netherlands
May 2019

Foreword

This thesis has been written at Tilburg University with the great help of my supervisor dr. Ç. Güven. Going into this subject I had relatively little knowledge about the influence of education on economic prosperity. However, researching and writing this thesis has made me more aware of the possibilities education has to offer.

Maud Calame,
Eindhoven, May 2019

Abstract

The goal of this research is to create insights in whether historical data can help us to predict upwards and downwards trends in economic growth based on Education Development Indicators (EDI's). In order to find a classification model that best fits the data to predict these trend movements a Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF) and, a Support Vector Machine (SVM) were established and compared based on their performance. It is the aim of this study to supplement the lack of future insights in terms of economic growth predictions due to EDI's. In order to reach the goal of this study a research question (*by looking at historical data, can we accurately predict GDP movement trends based on access, spending, and years of education?)* and a sub question (*What classification algorithm performs best in distinguishing the up and down trends in economic growth based on access, spending, and years of education?)* were formulated. The results to these were as followed; for the sub question: out of the 4 classifiers, the outcomes showed that the Linear SVM performs best on the unseen data. These outcomes give the possibility to answer the overall research question: the outcomes for all classifiers showed a much better outcome on the train set compared to the test set. The prediction outcome for our best performed classifier had an F1-score of 0.22. This cannot be seen as an trustworthy representation.

Keywords: Economic Growth Prediction, GDP, Binary-classification, Logistic Regression, K-Nearest neighbor, Random Forest, Support Vector Machines

Table of contents

1. Introduction

1.1. Goal of research

The goal of this study is to create insights in whether historical data can help us to predict upwards and downwards trends in economic growth based on Indicators of Education Development. In order to find a classification model that best fits the data to predict these trend movements a Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF) and, a Support Vector Machine (SVM) will be established and compared based on their performance. It is the aim of this study to supplement the lack of future insights in terms of economic growth predictions due to EDI's.

1.2. Context of thesis

Education is a key-word in the field of economic growth. This is because when speaking of economic growth, education is seen as one of the tools that can be used to accomplish this growth. Academia has long showed an interest in the relation between education and economic growth. An example of this is the endogenous growth theory. This theory hypothesizes that human capital is a main contributor in economic growth (Lucas, 1988; Romer, 1990). Human capital is well-defined by the OECD as: "the productive wealth embodied in labour, skills and knowledge" (OECD, 2001). Looking at more recent studies, the increase of educated people has been emphasized as a key contributor of economic growth (Gradstein and Justman, 2002; Hanushek and Woessmann, 2012; Pelinescu, 2015). Education as a tool for economic growth has also been an often-visited topic by famous politician and public figures. Nelson Mandela, anti-apartheid activist and politician stated the following: "education is the most powerful weapon which you can use to change the world" (Mandela, 2003). Mandela drew this knowledge from earlier work by Schultz (1961): The Human Capital Theory (HCT). In this theory Schultz argues that investment in people's education increases their productivity and thus leads to economic growth. Likewise, Rieckmann, Mindt and, Gardiner (2017) and Wals and Benavot (2017) highlighted education as an essential requirement to achieve sustainable development for societies across the world. The argument that all of these scholars share is that each individual needs a certain level of skills, knowledge, thoughts, and beliefs to become an effective individual for our society and achieve educational development (UNESCO, 2014).

This thesis aims to predict movements in economic growth based on EDI's using historical data. Therefore, the field of economic growth is part of the research domain. Within that domain this thesis will focus in particular on economic growth due to educational development. These domains and themes are of increased importance and interest because of our knowledge intensive society. In the Western world the term knowledge economy is often used, and the quality of education is overall high. However, this is not the case everywhere. The lack of qualitative education remains in the current top ten main global issues facing the world (United Nations, n.d.). Despite the fact that the relationship between economic growth and education has been studied widespread (this issue will be addressed extensively in the section related work), little attention has been paid to the ability to predict economic growth based on EDI's by using classification algorithms. The fairly new field of big data offers researchers a new and unique approach to the material. By combining the state-of-the-art methods with the importance and relevance of the topic, makes this thesis stimulating.

1.3. Scientific relevance

Trusting and making more accurate economic growth predictions based on EDI's can contribute to useful insights for appropriate education policy making. This then in turn can help make more accurate strategies for resource allocation and early decision-making. The positive relation between investment in education and economic growth is widely studied (Schultz, 1962; Becker, 1962). However, too little attention has been paid to predicting economic growth trends based on exploratory EDI's. The EDI's that are of interest in this study are related to access and participation in different levels of education, spending on education by the government, and years of schooling (for short, access, spending, and years of schooling). It is the aim of this study to supplement the lack of future insights in terms of economic growth predictions due to EDI's. The reason this has scientific relevance is because making accurate future predictions by using data models has high potential but also high consequences. If performed correctly, predictions can be very accurate and used to help the economic progress of countries. However, if performed incorrectly people can make predictions based on inaccurate data. Therefore, academic additions to this field are of vast value.

1.4. Practical/societal relevance

Next to scientific relevance this thesis also has a practical and societal value. Studies underline the positive effect of investment on education on the productivity of individuals. Productivity can increase because citizens can use their skills and knowledge to complete tasks more efficiently (Becker, 1962; Kampelmann, Rycx, Saks and Tojerow, 2018). Furthermore, having a higher productivity rate provides the opportunity to offer and sell more goods and services with the same input, which then in turn leads to a higher GDP. The outcomes of this study can help frame a successful strategy to achieve this goal for countries and societies. However, improving quality is not always easy because governments have limited financial resources to spend on education. Therefore, the important and challenging task that arises from this, is that governments have to spend their limited financial resources as effectively as possible. The outcomes of this thesis can provide policymakers and decisionmakers with information about the impact of the EDI's (access, spending and years of schooling) on economic growth. If these indicators show good prediction results for economic growth, there can be a scope for revisiting governments financial budgets. Then more attention can be payed globally to solve the problem of societal inequalities due to educational concerns. Better resource allocation due to effective policymaking can lead to more educated people which contribute to higher productivity in a country and thus to economic growth. The practical and societal relevance of this thesis is that its results can contribute to better resource allocation, more educated people, and in turn higher economic growth.

1.5. Problem statement and research question(s)

The right to education has been adopted by the Universal Declaration on Human Rights since 1948 (UN Assembly, 1948). Nevertheless, this human right is still unreachable for millions of children around the globe due to lack of (financial) resources, access to good education, and lack of cognitive skills (Hanushek and Woessmann, 2012). As stated in the beginning of this chapter: the goal of this study is to create insights in whether historical data can help us to predict upwards and downwards trends in economic growth based on education development indicators. Based on this goal the following problem statement (PS) has been formulated.

- *PS: To what degree can a classification algorithm distinguish upwards and downwards trends in GDP by looking at Indicators of Educational Development?*

In order to find an answer to the PS, this study will address the PS with 1 main research questions and 1 sub research question. The first research question is:

- *RQ1: By looking at historical data, can we accurately predict GDP movement trends based on access, spending, and years of education?*

This study examines whether we can predict if GDP goes up or down by looking at the exploratory variables related to access, spending, and years of education with classification. To find an answer to RQ1 the following sub question needs to be answered:

- *RQ1.1: What classification algorithm performs best in distinguishing the up and down trends in economic growth based on access, spending, and years of education?*

This problem is seen as a classification problem, because predicting the actual level of GDP is a much more complex task since GDP depends on many influential factors. To answer RQ1.1 a LR, RF, KNN, and a SVM classifier will be investigated in order to find the model that best fits the historical data to conduct the GDP trend prediction. The dependent variable is GDP (up/down) and the independent variables are (enrolment in secondary education, enrolment in tertiary education, government expenditure as % of GDP, and expected years of schooling).

1.6. Findings

After comparing the performance of the four classifiers on the unseen data, the results showed that the Linear SVM classifier performs best on the historical data. Unfortunately, the results on the train set were much better than on the unseen dataset that does not give the ability to generalize well. A reason could be that the economy is a complex entity that depends on many factors. Furthermore, countries (high income versus low income) differ in their (economic) behavior which can make it hard to come up with a single relationship that works for all countries to achieve economic properity.

2. Related work

In this section, the context of the topic will be explained in section 2.1. In section 2.2. relevant work will be described, and in section 2.3. research gaps and shortcomings will be presented.

2.1. Area of research

2.1.1. Provide context

The endogenous growth theory was introduced in the previous chapter of this thesis. Since it has had an important influence on the field, many studies have focused on the impact of human capital on economic growth. This research aims to contribute to this field, taking the approach of big data. The problem that will be addressed in this thesis is "*to what degree can a classification algorithm distinguish upwards and downwards trends in GDP by looking at EDI's?"*. An answer to this problem statement will help to fill in the gaps in the existing literature. In the past many works have mainly been quantitative, in the sense that they were conducted before the era of big data and thus did not use the possibilities this method provides. In addition to this, not much peer reviewed literature can be found on predicting the direction of GDP and certainly not on the basis of EDI's. On top of that, this study will try to make binary distinctions between upwards and downwards trends of GDP by testing several classification algorithms.

2.1.2. Research issue this thesis focuses on

This research is conducted in order to test if GDP trend predictions can be made based on EDI's. The goal of this study is to find a prediction model that best fits the data by comparing their prediction accuracy. A LR, KNN, RF, and SVM classifier algorithm will be executed. The **independent variables** are enrolment rates in secondary education, enrolment rates in tertiary education, government expenditure (as % of GDP), and expected years of schooling. To make predictions of upwards and downwards trends, GDP per capita (Purchasing Power Parity) is used as a reference point for the **dependent variable:** GDP up and down. GDP Purchasing Power Parity (PPP) takes into account the variations in the exchange rate which are valuable when comparing countries from different areas around the world. The dependent variable used is GDP per capita (PPP) but will be referred to as GDP in this thesis. An important result of accurate economic growth predictions (based on EDI's) can be to achieve a better decision-making process, allocation of government budget resources, and monetary policymaking in the field of education. Especially, more budget and efficient resource allocation can then in turn lead to more educated people in a country which can lead to more successful achievements and growth opportunities as mentioned in the introduction section.

2.1.3. Why is it important

Because governments have limited budgets to spend on education, exploring the functional relationship between the EDI's and GDP is an important contribution for researchers, policymakers, and decisionmakers (Delurgio, 1998). In that sense, this thesis aims to provide models to assess the degree of influence of certain EDI's on economic growth and to aide policy makers to improve budget allocation for education. Which then leads to raising the level of education in countries and with higher levels of educated people, a raise in economic growth can be achieved.

2.2 Relevant work in the same research area

2.2.1. Relevant theories

The American economist Theodore Schultz started promoting education in a form of human capital in 1961, which would contribute to the economic growth of countries worldwide. He argued that government spending on education in a form of skills and knowledge development leads to an increase in the quality of human effort and its enhanced productivity (Schultz, 1961). The importance of investment in human capital has also been supported by Romer (1986) who argued that an investment in human capital will lead to national economic growth. Furthermore, Ozturk (2001) assumed that countries cannot secure desirable human, social, and economic growth without government expenditure on human capital. Throughout Western education systems, Schultz's HTC is seen as an important influence of economic success (Fitzsimons, 2017). What we know about the impact of education on economic growth is largely based upon growth theories that link productivity to the transmission of knowledge and to human capital (Nelson and Phelps, 1966; Mankiw, Romer and Weil, 1992). As stated before, Schultz (1961) concluded that investment in people's education increases their productivity and thus leads to economic growth. Additionally, Barro (1991), Barro and Lee (1994), and Psacharopoulos and Patrinos (2018) also found positive effects of education on economic growth. Hanouz and Khatib (2010) have researched the relation between education and productivity in Arab countries. The conclusion of their study was that in order to achieve higher productivity levels, the quality of education that people receive needs to be better. Likewise, several studies describe education as an essential requirement to achieve sustainable development (Rieckmann et al., 2017; Wals and Benavot, 2017). The researches mentioned thus far provide evidence that educational development has a positive impact on economic growth.

A logical next step in the field of education as a role in economic growth, was economic growth prediction. A considerable amount of literature has been published on economic growth prediction. In line with our field of research, many studies have examined the positive relationship between education and economic growth (Barro, 1991; Bils and Klenow, 2000). To measure a country's economic well-being the Gross Domestic Product (GDP) is a worldwide used quota. GDP is defined as: "a market value of all final goods and services produced within a country in a given period of time" (Mankiw and Taylor, 2006, p. 468). Governments, policymakers, and businesses use economic growth or GDP as a reference point for decision-making, strategies for allocation of government budgeted resources and monetary policy (Montgomery, Jennings and Kulahci, 2016). Therefore this thesis will also use GDP as an indicator of economic growth. Since data from different countries is used, the decision has been made to use GDP per capita (PPP) which allows for the comparison of countries with different (population) sizes and variations in exchange rates. More insights about the effect of EDI's on economic growth trends, can result in higher priority for good education in countries and better allocation of government budgeted resources towards the education sector (De Witte and Lopez-Torres, 2017). GDP is a good measurement for economic growth. However, not much peer reviewed literature can be found on predicting the direction of GDP (with the help of machine learning methods) and certainly not on the basis of EDI's (quantitative).

Nonetheless, there has been some recent research into big data economic growth prediction. Junoh (2004) investigated the possibility of GDP predictions based on knowledge-based-economy indicators by using a neural network algorithm. In Junoh's study the results of the neural network algorithms were compared to an econometrical model which resulted in a better outcome of the machine learning algorithm. This research showed that there is possibility for predicting economic growth with machine learning algorithms to solve academic research problems. Junoh focused on ICT variables to predict GDP. This can be valuable for current research because it provides evidence that successful GDP predictions can also be made in the education sector. Likewise, many policymaking and decision-making is vastly based on economic factors such as GDP.

Nowak and Dahal (2016) explored the contribution of educational levels on economic growth. Their Ordinary Least Square model was used to estimate three parameters: primary, secondary, and tertiary school. In primary school children learn the basic skills needed such as mathematics, reading and writing in order to continue to secondary school in which more subject-oriented skills will be learned. In addition to this, tertiary education is mainly focused to develop specialized qualification to contribute to the labor market (UNESCO, 2012). The results of Nowak and Dahal's study showed that all three parameters influence economic growth. Their three parameters are of importance to this study to track which level of education has the highest contribution to economic growth. This research can serve as guidelines for policymakers and decision-makers to spend their budgets more accurate. Next to Nowak's and Dahal’s parameters that influence economic growth, Marković, Petković, Nikolić, Milovančević, and Petković (2017) tried to make economic growth predictions based on scientific and technological indicators. These are indirectly related to education, because they signify a level of education maturity. In their study the accuracy of an Extreme Learning Machine (ELM) algorithm was compared to genetic programming (GP), Artificial Neural Networks (ANN) algorithm and fuzzy logic algorithm outcomes. The goal of this study was to create an ELM algorithm that can accurately predict economic growth. The results showed that from the algorithms tested, the ELM performs best on the data. Which gives evidence that GDP predictions can be made by using machine learning algorithms.

Returning to GDP; Ermişoğlu, Akçelik, and Oduncu (2013) made predictions of GDP growth by using credit data. Their study was not related to the effect of education on economic growth, but is relevant because the fact that GDP growth predictions are made based on data from different sectors underlines the importance of good GDP growth predictions. One of the main reasons why credit data was used is because of the access to this data. Credit data is available on a daily basis whereas GDP data is made available on a much less frequent base. However, the outcomes showed that the credit data gives positive results for both nowcasting as forecasting of GDP movements. Next to this, Richardson, Mulder and, Vehbi (2018) compared statistical forecasting techniques and machine learning techniques to predict GDP growth, which results in most accurate performance for predictions using machine learning algorithms. These outcomes imply that using machine learning algorithms can lead to higher accuracy when performing a GDP growth prediction task. Therefore, this will be applied in the current study in which we try to predict upwards and downward trends in GDP based on EDI’s by testing well-known classification algorithms on their performance.

Most prior studies have used general statistics instead of machine learning methods to find correlations between independent variable(s) and the outcome variable. Bang, Sen and Basuchoudhary (2017) mentioned an important advantage of machine learning algorithms. They argued in their research that concerns of cross-validation and out-of-sampling predictability should be taken into account. This means, that this research will evaluate the learning outcomes of the algorithms for accuracy with unseen test data. Bang, Sen and Basuchoudhary (2017) then argued that this can be important evidence for policymakers, because building their policies based on outcomes that are not tested on new real-world situations can raise questions about the reliability on which policymakers make their choices.

Despite the fact that many researches have underlined the positive relationship between education and economic growth, and the ability to make GDP predictions based on a variety of indicators as discussed extensively in the introduction and related work section. There are still some controversial outcomes that need to be discussed.

2.3 Research gaps and shortcomings

2.3.1. What is missing from prior research

A closer look into the literature on economic growth prediction, however, reveals a number of gaps and shortcomings. Most studies have relied on predicting economic growth with data from other sectors, such as ICT and financial indicators. Next to this, according to the literature predicting a variable's direction by using machine learning techniques is mostly done in the stock market field (Patel, Shah, Thakkar, and Kotecha, 2015; Ballings, Van den Poel, Hespeels, and Gryp, 2015; Basak, Kar, Saha, Khaidem, and Dey, 2019). Furthermore, previous mentioned studies were mostly conducted by using more complex machine learning algorithms, and are therefore hard to recreate. Nevertheless, since there is no clear evidence that complex models always perform better in their predictive or explanatory power than simple models in this study, both were performed. To fill this gap in the literature, this study will use EDI's to predict the direction of GDP by focusing on three main contributors, namely access, spending, and years of schooling as mentioned in section 1.5. For the problem statement and research question(s) please refer back to section 1.5.

2.3.2. Limitations of existing models

The above-mentioned models and theories add to the existing framework of academic work. Furthermore, some literature is found on economic growth predictions. There are however, limitations to those models. For instance, Feng and Zhang (2014) who developed an artificial neural network algorithm for economic growth prediction. Their algorithm was vastly more complex than current research, and hard to recreate. Nevertheless, more literature is written in other fields such as stock market movement prediction by using machine learning algorithms instead of economic GDP predictions as mentioned before. These studies show that machine learning tasks can solve classification problems in other fields which give potential evidence for successful predictions in this research. However, it is of importance that the researches can be recreated and applied on different data sets.

2.3.3. How this thesis will fill the research gaps

This research focuses on economic growth prediction facilitated by education. More insights about the impact of potential valuable educational determinants of economic growth can result in better decision-making, allocation of government budget resources, and monetary policymaking (De Witte and Lopez-Torres, 2017). Furthermore, as previously discussed, an effective delivery of educational resources to citizens can lead to a higher amount of educated people, and this then in turn can lead to better and more equal societies.

The goal of this study is to create insights in whether historical data can help us to predict upwards and downwards trends in economic growth based on EDI’s. This problem needs to be addressed because it is important for the direction of the GDP trend serves as an influential reference point for decisionmakers to determine the amount of government budget resources to spend on a specific sector. In addition to this, accurate predictions of GDP trends based on EDI’s can help governments to reduce inadequate spending and improve spending governmental budgets much more efficiently.

Globally comparable EDI’s are used for the prediction the GDP direction (World Bank, 2018; UNDP, 2018). The independent variables used are: (1) access to education measured by enrolment rate in secondary school, (2) enrolment rate in tertiary school, (3) spending on education measured by government expenditure as % of GDP, and (4) expected years of schooling. In this study the dependent variable is GDP per capita (PPP). These indicators have been chosen to test if these specific EDI’s can perform accurately predictions of upwards and downwards trends in economic growth. Since governments have limited budgets to spend on education, creating insights into functional links between the EDI’s and GDP is an important contribution for policymakers and decisionmakers. Delurgio (1998) refers to functional links as exploring this relationship between independent variables and the dependent variables because it can support researchers to determine if there is any uncertainty between the variables that are used.

To the best of the author’s knowledge, economic growth trend predictions based on access, spending, and years of schooling by using the classification algorithms LR, KNN, RF, and SVM have not been explored yet. Besides, little attention has been paid to test models that can accurately predict economic growth based on historical data of EDI’s.

5. Methods & Experimental Setup

Finding the most suitable algorithm for a prediction task is not a straightforward process. Trying out several algorithms in order to find the most accurate outcomes is a trial and error process. Wolpert and Macready (1997) described in their study the “no free lunch theorems for optimizing models”, which can be interpreted as that finding a well performing algorithm is an ambiguous process until you test several algorithms to find out which one performs the best.

3.1. Dataset description

In this thesis datasets are combined from the World Bank (WB) and the United Nations Development Programme (UNDP). The datasets are based on EDI’s for different countries all over the world.

3.1.1. World Bank (WB)

The WB has a specific database for education topics, called EdStats (Education Statistics), this includes World Development Indicators related to education. The data made available by the WB came from officially-recognized (these include official country data and multinational organizations) institutions. Data from 1999 till 2015 has been collected. This timeline was chosen based on two criteria. First, because many data points before 1997 and after 2017 were missing. Too much missing data can potentially bias the outcomes that results in a misleading reflection of the real situation. Second, because the data from the UNDP described in section 3.1.2. only collected data from 1997 till 2017. The Edstats database includes country-specific educational information, based on yearly time-series from 1960 till 2018 for more than 4000 internationally comparable indicators from 217 countries worldwide. Since one of the main challenges in educational data is the many missing values, the WB has selected a group of 7 EDI’s that are relevant to improve the economic well-being of a country as well as a relatively well filled dataset. These indicators are collected for this study as potential independent variables to choose from in order to predict the upward and downward trend in GDP.

3.1.2. United Nations Development Programme (UNDP)

The UNPD created a database filled with human development data. In this study the data that belonged to the education dimension has been extracted as potential predictors. The education data consists of 25 indicators for 189 countries, based on yearly time series from 1997 till 2017. Due to many missing data points in some indicators, and overlap with the World Bank data selection of variables some data entries have been excluded from analysis.

3.2. Variables

3.2.1. GDP per capita (Purchasing Power Parity)

As defined in the related work section, GDP can be seen as the value of all goods and services produced within a country. Because countries differ in many aspects such as standard of living, demographic size, population size and more. The GDP per capita that takes into account the population size of a country is used. This makes it possible to compare countries to each other. Next to this, in this thesis the GDP per capita (PPP) indicator is used as a reference point to classify the upward and downward trends in GDP. The reason is that this indicator takes into account the cost of living that differs from country to country.

3.2.2. Exploratory variables

Since economic growth depends on many factors in different areas, it can create a great opportunity for researchers to investigate determinants that have a significant impact on the economic growth to make better projections for the future. Big data gives researchers the opportunity to measure, research, and discover reasons for economic growth. However, we are still at the base of creating this new field. Especially with the rise of knowledge-based economies, more research has to be done into using big data to analyze and further help our knowledge intensive societies. This is because big data can provide insights and further policy-making in this area.

In developing countries (especially rural areas), the level of access to education is improving but still has to deal with problems such as bad educational systems, too little teaching materials, and educated teachers (Vasconcellos, 1997). According to UNESCO (2018), in low-income countries there are 21.9 million children that drop out primary school (20.3%) compared to 2.6 million (3.5%) in high-income countries. This implies that access to education is a possible predictor for economic growth and therefore important for policymaking and decision making (UNESCO, 2018). Lutz, Creso Cuaresma, and Sanderson (2008) stated that access and participation in certain levels of education are important as predictors of economic growth since it contributes not only to better incomes for individuals, but it also serves as a prerequisite for economic expansion. In this study, enrolment rates for secondary school and tertiary school were chosen for the analysis. This has to do with the overall high enrolment rates in primary education due to the goal of the UN for universal primary education. There are much more interesting differences in access and partitioning in secondary and tertiary education. For this reason, we focus on these indicators. Furthermore, lack of financial resources can also be an important reason why children do not attend school (Hanushek and Woessmann, 2012). Therefore, more expenditure on education by the government can contribute to a declining financial burden for parents with low incomes. This then in turn can lead to higher enrollment rates in school. This suggests that government expenditure can be a valuable predictor for GDP trend prediction. Additionally, Appiah (2017) find that more government expenditure on education positively affect GDP per capita in developing countries. Moreover, Hanushek (2016) stated an important aspect for policymaking in terms of relation between the number of years of schooling and economic growth. Their study argued that knowledge learned in the past is necessary to expand learning skills in the feature. Thus, in the end, more people with a certain level of skills, knowledge, thoughts, and believes can result in enhanced productivity this then in turn will lead to upward trends in GDP. Therefore, years of schooling can be a useful indicator to predict upwards and downward trends in economic growth.

3.3. Preprocessing

For this study, data accessible from both data sources in the period of 1997 till 2017 were examined. Since the information for every one of the three datasets were covered for period, these specific years were chosen to guarantee consistency. However, due to relatively high missing data for the years 1997, 1998, 2016 and 2017, the decision was made to use only the years from 1999 till 2015 for the analysis. Furthermore, the countries represented at the WB and UNDP are somewhat different. For this reason, only those countries that were available in both datasets were chosen for this study. Moreover, there were some countries that were in both datasets but had a slightly different name, these were made equal so that no countries were removed for which information is available in both datasets. Since both datasets are constructed by using information from officially-recognized sources there is some variable overlap. In most cases these variables only have a slightly different variable name. these variables have been compared for consistency and dropped if there was a complete overlap. After this, 24 potentially indicators were left. To keep this research specific and concise, only variables related to access, spending ,and years of schooling were chosen for analysis. Next to this, because missing data was a problem for educational dataset, the decision has been made to choose variables with not too much missing data. This is since too much missing data can bias the results too much. All potentially indicators were merged into a data frame and a selection was made for the analysis. Before merging all data frames, the data was reshaped from a wide format to a long for binary classification.

From the WB the following EDI's were selected for the analysis per country, based on a yearly time series from 1999 till 2015:

- Enrolment in secondary school
- Enrolment in tertiary school
- Government expenditure as (% of GDP)

These potential predictors were used in assessing if these upwards and downward trends in GDP can be predicted based on these EDI's. Next to this, the GDP per capita (PPP) variable is also available via the WB and was used as a reference point to classify data for a specific country if the GDP goes up or down for this country in comparison with last year. From the UNDP, to assess if the time that individuals spend in school influence the economic well-being of a country, the following EDI was selected for the analysis per country, based on a yearly time series from 1999 till 2015:

- Expected years of schooling

In total, 39 countries were selected for the final analysis, since these countries had a maximum of 2 missing values per prediction variable for a country over the years 1999 till 2015. This was done, because imputation of many missing values can lead to highly biased results, which is not desirable. To conclude, the dataset used for the classification task consisted of 663 observations from 39 countries over a period of 17 years.

3.3.1. Data imputation (Amelia II) & Limitation of Amelia II

Despite the fact that the data was selected based on the criteria of maximum of 2 missing data points per prediction variable for a country over the years 1999 till 2015 there was still data missing. Honaker, King, and Blackwell (2011) came up with a method for handling missing values in time-series cross-sectional data: Amelia II which is available in R. One of their assumptions stated for the Amelia II package is that it assumes that the dataset is multivariate normally distributed. In the real world this is often not the case. Plotting histograms showed that variables were different form normally distributed. However, evidence was found that the Amelia II method functions just as other more complex models (Honaker, King, and Blackwell 2018). Therefore, the Amelia II package was used to fill the missing data points which gives the possibility to construct analysis without the problem of missing data.

3.3.2. Feature normalization

The prediction variables will be normalized before modeling. First, the data will be centered, followed by subtracting the mean and, finally dividing by the variance to normalize it. The ranges will be equalized so that every feature gets approximately the same weights which leads to more numerical stability (Aksoy & Haralick, 2001).

3.3.3. Data partitioning

Since a 80/20 data split is quite a commonly occurring ratio, the data was partitioned into a training set of 80% of the data and a test set of 20% of the data. The trained models will be compared based on their performance on the training data by using 10-fold cross-validation.

This cross-validation method allows to use the maximum available data in the process of training and testing the model, since the data sample is somewhat limited due to imbalanced classes. In the end, after parameter tuning the test set will be used to assess the performance of the algorithms. This is done to check the capacity of the algorithm to make predictions on data it has never seen before.

3.3.4. Imbalanced data

In this study there is a highly uneven class distribution which can lead to an accuracy paradox. This means that the model will be biased towards the majority class which can lead to unrealistic high accuracy (Chawla, 2009). The Dyplr package available in R allows for the calculation of the ratio between the times GDP goes up and down. The outcomes showed that the educational dataset is highly imbalanced (ratio of 11:89) which needed extra attention to prevent for biased results. This is because in 89% of the cases the GDP per capita (PPP) followed an upwards trend. For this reason the decision has been made to apply the SMOTE method to the trainset. This means that an equal percentage of up and down observations were used for analysis in the train dataset.

Bellinger, Sharma et Japkowicz (2012) tested the performance of binary and one-class classifiers on imbalanced datasets in which the results showed that imbalanced datasets had a negative effect on the performance of a binary classifier. Therefore, the SMOTE method was used as a resampling approach (Chawla, Bowyer, Hall and Kegelmeyer, 2002). New instances belonging to the minority class will be generated based on the nearest neighbors of these instances (Chawla et. al., 2002). The advantage of the SMOTE method relative to simple oversampling and under-sampling methods is that it not simply makes duplicated of instances from a specific class which can result in overfitting or removes instances from the majority class which can lead to the deletion of important information for modeling. But it generates instances based on a convex mixture of neighboring cases. SMOTE is available via R package DMwR and is a well-known method to handle classification problems that have to deal with imbalanced datasets. This method gives the opportunity to learn from imbalanced data, by making the dataset balanced which will avoid biased results regarding the majority class.

3.4. Experimental procedure

In this study the following classification problem will be examined: "*by looking at historical data, can we accurately predict GDP movement trends based on access, spending, and years of education?*". In order to find the model that best fits four classifiers will be tested (LR, RF, KNN, and SVM). This is done using predictors related to access, spending and years of education.

3.4.1. Algorithms

This study aims to help fill in the lack of insights in economic growth trend prediction based on these EDI's by testing well-known classification algorithms to find a model that best fits the historical data.

3.4.1.1. Logistic Regression algorithm

Logistic regression (LR) is an algorithm commonly used to conduct (binary) classification tasks. The goal of this classifier is to determine to which class a particular instance belongs. This study will use binary classification to predict if the GDP goes up or down. Chin, Geweke, and Miller (2000) conducted some similar work in which they predict turning points of unemployment. Their study underlined that a binary classification was conducted in which GDP was classified according to growth or decline before the LR was carried out. For this thesis, we will create a new variable with a binary distinction between upward and downward trends in economic growth as the dependent variable. A LR model has been used to predict class membership (GDP up or down) based on EDI's:

1) $p(X) = \frac{\mathrm{e\beta 0+\beta 1X1+\cdots+\beta pXp}}{1 + \mathrm{e\beta 0+\beta 1X1+\cdots+\beta pXp}}$

Where P(X), is the probability that upwards and downwards trends can be distinguished based on the EDI's in the year after. $X = (X1, \ldots, Xp)$ are the EDI's (predictors), e is the base of the natural logarithm, the coefficients attached to the independent variables β0 (intercept) and β1 + … + βp (slopes) are unspecified, and will be estimated via the training data in R.

3.4.1.2. Random Forest algorithm

A second classifier that will be trained for upwards and downwards trends in economic growth is the Random Forest (RF) algorithm which can be used for both classification and regression (Breiman, 2001). Breiman highlighted the conflict between the simplicity and accuracy in models. One of the main conclusions was that a decision-tree algorithm is generally easier to interpret then a RF algorithm. However, a RF algorithm usually outperforms a decision-tree based on accuracy. Because this study focuses on the model that best fits based on accurate performance a RF classifier will be tested. A RF algorithm creates multiple decision-trees and grows due to a random component that is introduced to each decision-tree separately (Breiman, 2001).

The RF classifier is formulated and explained by Breiman (2001) as:

2) $\{h(x, \Theta\mathrm{k}),\ \mathrm{k} = 1, \ldots, K\}$

Where, $h(x, \Theta\mathrm{k})$ represents a single tree predictor, $(\Theta\mathrm{k})$, a random vector (independent and identically distributed), $\mathrm{k} = 1, \ldots, \mathrm{K}$, amount of trees produced, and x = input vector that will be classified by using the majority votes from the single tree predictors.

3.4.1.3. K-Nearest Neighbor algorithm

The third algorithm that will be discussed is the K-nearest Neighbor (KNN) which also can be used for both classification and regression. This classifier uses a distance metric to find the nearest neighbor instances (K) from a single test data point. A commonly used metric is the Euclidean distance and, can be mathematically formulated as[1]:

$$3)\quad d(x, {}'x) = \sqrt{(x_1 - x'_1) + (x_2 - x'_2) + \cdots + (x_n - x'_n)}$$

Where x and ‘x are different data points from which the distance is calculated. The optimal value of K will be found automatically by using cross-validation when training the model in R. Not much evidence has been found for the use of a KNN algorithm for predicting upwards and downward trends in economic growth. Nevertheless, much evidence is found in other sectors such stock price direction prediction (Ballings et al., 2015), predicting if a patient gets a heart attack (yes/no), currency exchange rate and more (Imandoust and Bolandraftar, 2013).

3.4.1.4. Support Vector Machine

The last algorithm is the Linear SVM, which can also be used for regression and classification. This model uses hyperplanes to separate classes in the data. The “kernel trick” is a phenomenon in machine learning that helps a linear classifier to perform a classification task on non-linear data by setting it into a higher dimension (Mountrakis, Im, and Ogole, 2011). Finally, (Boser, Guyon, and Vapnik, 1992) stated that this classifier performs well on a limited amount of training data which leads to the ability to generalize well on the test data. The linear SVM classifier is formulated by James, Witten, Hastie, and Tibshirani (2013, p. 351) as follows:

$$4)\quad f(x) = \beta 0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

The linear function of x is notated as$f(x)$, where n is the number of parameters a_i with $i = 1, \ldots, n$. The coefficients for $\beta 0$ *(intercept),* and a_i (parameters) are calculated with the inner product $\langle x, x_i \rangle$ between the new instance x and the existing instances in the dataset x_i. Nevertheless, a detailed explanation of the inner-product is beyond the scope of this thesis. It can be automatically done by R using the ‘linear’ kernel trick.

[1] There are multiple distance matrices available. However, to properly explain the workings of the KNN model, the matrices are explained by the Euclidean distance.

3.4.2. Implementation

The R software package version 3.4.1 was used for this study. The datasets as described in section 3.1, will be extracted as a csv file and will be loaded into the development interface of R, called R studio (version 1.0.153). The following R packages were used for this study:

- Readr package to read several csv files into R.
- Dplyr package for manipulating the data.
- Tidyr package to make the original wide dataset long with the gather() function. This is done because the columns in the original datasets are yearly time-series and not variables.
- Base package is used to merge data frames.
- Ggplot2 package has been used for data visualization.
- The Caret package is used for the setup of the prediction model in which the e1071 package supports the Caret package, and set.seed() for reproducible.
- Amelia II, extended by the Zelig package for handling missing values in time-series cross-sectional data.
- DMwR package for the SMOTE used as a resampling approach to work with imbalanced datasets.

3.4.3. Evaluation criteria

For the classification problem of upwards and downwards GDP trends prediction a confusion matrix will be used to evaluate the performance of the models. The accuracy is the most well know measure for classification problems. Nevertheless, when the class probabilities differ too much, this evaluation metric can be misleading. The F1 score can be seen as a harmonic mean between precision and recall, and can be used to evaluate the model performance when dealing with imbalanced classes (Tharwat, 2018). Next to this, Cross-validation is used to evaluate the model on the train set since it might be the case that the model performs well on the available data but not on new data which is problematic for performance generalization (Refaeilzadeh, Tang & Liu, 2009). The 10-fold cross-validation method used in this study allows to use the maximum available data in the process of training and testing the model, since the data sample is somewhat limited due to imbalanced classes. In appendix I, detailed information about the confusion matrix and the calculations for the evaluation metrics can be found.

4. Results

4.1. Best performance model

This study examines whether we can predict if GDP goes up or down by looking at the exploratory variables related to access, spending, and years of education with classification. Observations were classified as up or down using several supervised machine-learning techniques. To assess which classifier best performs on the classification task, various models have been tested. In order to answer the research question the approach with the highest F1 score needed to be found. The problem that can arise is that too little samples of the minority class make it more difficult for the model to learn patterns that distinguish classes from each other. Due to highly imbalanced data the SMOTE method was applied on the training dataset. Furthermore, before splitting data into train and test set, feature normalization was applied on predictor variables (for more information please refer back to section 3.3.2.).

The next step in the experiment was to optimize the parameters with parameter tuning. In Table 1, the best parameters are shown. 10 fold-cross validation was applied in the parameter tuning process. It was used for evaluating the model performance on the train data.

Table 1
Parameter tuning and its F1 scores by using 10-fold CV

	F1 score (average out of 10 CV samples)	
Classification models	F1	Best parameters
Logistic Regression	0.62	
Random Forest	0.81	‘mtry’ = 4
K-Nearest Neighbor	0.81	‘K’ = 1
Support-Vector Machine	0.58	‘’cost’ = 1.75

Table 1: parameter explanation: mtry = the amount of variables to split on per tree node, K = the number of nearest neighbors and, cost = the cost of the misclassification.

The results of the F1 score are averaged based on the 10 fold cross-validation as represented in table 1. These averaged F1 scores came out of the analysis after the classifiers were tuned to find the best circumstances for the model to show the best results. Based on the outcomes generated by using 10-fold cross-validation it seems that the KNN classifier and the RF classifier performed best on the binary classification task that was given. However, to see if this is also the case when using unseen data without the SMOTE method the F1 scores for each classifier were compared based to the test set. The outcomes of this are showed in figure 1.

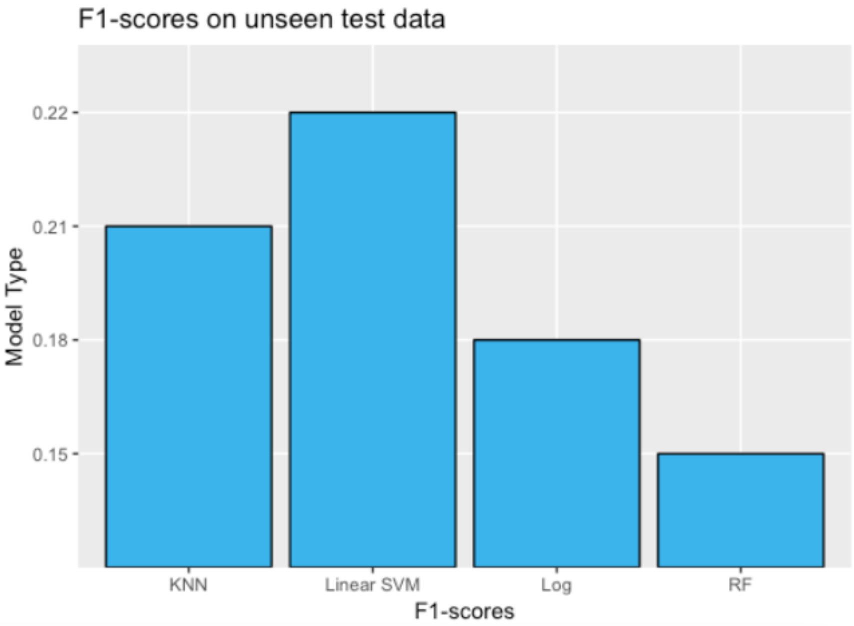

Figure 1: F1-scores on unseen test data

As shown in figure 1, the results on the test set are much less accurate than in the training set. From this figure, it can be concluded that our model has not generalized well to new, unseen data. The histogram in figure 1 showed no large performance differences between the classifier and the results are somewhat disappointing. Nevertheless, despite the small differences, it can be concluded that Linear SVM outperformed the other classifiers and performed best on the test set.

These results are a contradiction to the results from the train data in which the RF classifier and KNN classifier seem to perform best. This will be further explored in section 5.

Table 2 and 3 explain the performance outcomes of the Linear SVM classification model on the test dataset. In table 2 a confusion matrix shows these performance outcomes of the Linear SVM model on the test dataset. For more information about calculations of a confusion matrix see appendix I.

Confusion matrix		
	Reference	
Prediction, N = 132	**Down**	**Up**
Down	7	43
Up	7	75

Table 2: Confusion matrix, where N = total instances in the test set, reference = actual cases, and prediction = the predicted cases.

In total, 132 observations were used in the test set to evaluate the performance on the unseen dataset. From this, there were 14 cases that belonged to the “down” class, of these there were 7 cases that were correctly predicted and truly classified as going down. In total there were 118 cases of going “up", of which 75 cases that were truly going up were correctly classified as up. However, 43 cases that were going up were misclassified as down and 7 cases that were predicted as up but belonged in fact to be down.

Too specify, for the Linear SVM, 50% of the down samples were correctly classified and 64% of the up samples were correctly classified. The prediction of down samples was right only 14% of the time (precision), which means that in 86% of the cases the model was not able to correctly classify a down sample as down.

On the other hand, the prediction of the up sample was correct 91% of the time. This means that only 9% was not correctly predicted. Since we deal with highly imbalanced classes it could happen that actual “down” samples will be predicated as “up” samples. This can lead to misleading information for policy and decisionmakers. The down sample is stated as the positive class while the results show that from the actual down samples 50% was appointed to the up class whereas this is not the intention. This confirms with the first line in which was stated that 50% of the down samples were classified correctly. Because this research deals with unbalanced data which still can lead to relatively high accuracy the decision was made to measure the performance on the test set with the F1-score. The F1-score is the harmonic average of the precision and recall as mentioned earlier. In section 5, we will discuss the results in more detail. To conclude, despite the disappointing results, the Linear SVM performs best on the unseen data with a F1-score of 0.22 as presented in table 3.

Table 3 Outcomes on the test set for the Linear SVM classifier

	Accuracy	**Precision**	**Recall**	**F1**
	0.66	0.14	0.50	0.22

Table 3: Outcomes on the test set for the Linear SVM classifier

5. Discussion & Conclusion

The goal of this study was to create insights in whether historical data can help us to predict upwards and downwards trends in economic growth based on Education Development Indicators (EDI's). Below are the PS and the research questions:

- *PS: to what degree can a classification algorithm distinguish upwards and downwards trends in GDP by looking at EDI's?*
- *RQ1: By looking at historical data, can we accurately predict GDP movement trends based on access, spending, and years of education?*
- *RQ1.1: What classification algorithm performs best in distinguishing the up and down trends in economic growth based on access, spending, and years of education?*

This study uses different variables which can be found in section 1.5.[2]

Before training the models, the independent variables were normalized and the SMOTE method was applied. Next, 10-folds cross validation was used to tune the parameters. The parameters used were shown in table 1 in the result section. To evaluate which classifiers seem to be the best performers on the classification task to distinguish upwards and downwards trends in GDP based on the EDI's, the F1-scores were compared. Choosing the best metric for evaluating a specific classification task can be arguable since it depends on several aspects. In this study there was a highly uneven class distribution which could lead to biased results toward the majority class which can lead to unrealistic high accuracy (Chawla, 2009). Furthermore, the cross-validation method allows to use the maximum available data in the process of training and testing the model, since the data sample is somewhat limited due to imbalanced classes. It can give an idea about how well the model performs on unseen data without using the test set that is holding apart. The results in table 1 showed that the RF and KNN classifiers performed best on the training set. However, very contradicting results were found on the test set in which the performances were much lower compared to the train set and the Linear SVM showed the best results on the test set. Since these outcomes were unexpected the author explored reasons for these surprising outcomes. These are discussed below:

(1) The first problem may be caused by applying the SMOTE method. A first motivation would be that the train dataset reflected the distribution of the test dataset. Applying the SMOTE method before portioning the data into a train and a test set will give much

[2] This study examines whether we can predict if GDP goes up or down by looking at the exploratory variables related to (1) access, (2) spending, and (3) years of education with classification. The independent variables involved in the analysis were (A) enrollment in secondary school, (B) enrollment in tertiary school, (C) expected years of schooling, and (D) government expenditure. The GDP per capita (PPP) was used as a reference point to construct the GDP up or down factor variable for final analysis.

better results on the unknown data. However, since this method can create replicated instances from the classes a possible reason for high performance is that the instances represented in the training set will also be represented in the test set. This means that the test set cannot be presented as totally unknown data. For this reason, the SMOTE method is only applied on the training data by making the train dataset balanced. In this study the models were trained on balanced train data with the idea behind it that too few representative cases of the minority class (down) affect the model learning performance to learn patterns to distinguish "down" cases from "up" cases. Nevertheless, the results on the unknown dataset do not show evidence that the model really takes advantage of this. The cross-validation outcomes showed only impressively better F1-scores by using SMOTE for the RF and KNN classifier. In the other two classifiers no clear improvements were found. Even though the SMOTE method does not show better results for all classifiers it performs better or equal on the train set with approximately 60% fewer sample observations to work with. This is indicative on the positive effect of SMOTE on the predictive power. So, reduced cases available due to applying SMOTE was not problematic in that sense. Although, the SMOTE method was applied it seems that the model was not able to learn extensively from the extra minority cases to get good results on the unseen data. It can be concluded from this that our model has not generalized well to unseen data.

(2) Another reason for disappointing results can be the degree of impact of the prediction variables (access, spending, and years of schooling) on the movements in GDP. For example, Nowak and Dahal (2016) showed that access and portioning in different levels of education (primary, secondary, tertiary) positively influences GDP. However, the results imply that the combination of the prediction indicators related to access, spending, and years of schooling do not influence the GDP factor enough to assign correct classes on a frequent base.

(3) A third explanation can be overfitting. Overfitting refers to a model that models the training data too well. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize. Nonparametric models are more flexible (no assumptions) when learning a target function from the train dataset, which may have led to overfitting. To prevent this, it is a good way to pick out new models with more accurate data, and find better prediction variables for GDP movement.

Despite the disappointing results, the answer to RQ1.1 is as follows: out of the 4 classifiers, the outcomes showed that the Linear SVM performs best on the unseen data. These outcomes give the possibility to answer the overall RQ1: *By looking at historical data, can we accurately predict GDP movement trends based on access, spending, and years of* *education?* The outcomes for all classifiers showed a much better outcome on the train set compared to the test set. The prediction outcome for our best performed classifier had an F1-score of 0.22. This cannot be seen as an trustworthy representation. Referring this back to the PS this means that distinguish upwards and downwards trends in GDP by the selected EDI's do not give accurate results.

Referring back to the literature, many studies underlined the positive relationship between education and economic growth (Schultz, 1961; Becker, 1962; Romer, 1986; Barro, 1991; Bils and Klenow, 2000). Furthermore, prior studies conducted in other fields of interest showed evidence for successful GDP predictions (Junoh, 2004; Marković et al., 2017; Ermişoğlu, 2013). Moreover, Nowak and Dahal (2016) showed the positive contribution of educational levels on economic growth. Additionally, a lack of financial resources was stated by Hanushek and Woessmann (2012) as an important reason why children do not attend school, what makes government expenditure a valuable contributor for a higher school attendance rate. Likewise, Hanushek (2016) stated an important aspect for policymaking in terms of relation between the number of years of schooling and economic growth. Unfortunately, the theories does not support the data in this thesis.

5.1. Limitations

This thesis was a vast operation. However, as with almost every research this one has limitations. Below there are three reasons explored that had limiting effects on the study:

(1) It seems that the features are not powerful enough to predict GDP direction, using others may result in better outcomes.

(2) Besides the fact that the dataset was highly imbalanced, there were many missing data points in the cross-country data from the used sources. This makes the choice for predicting indicators very limited. What was striking was that many features were available, but the availability and trustworthiness of information fluctuated enormously per country. For this reason, certain choices were made, such as limiting missing values to a maximum of 2 in order to prevent the datasets from being made up of many artificially implementing values.

(3) This study used data from 39 countries with relatively little missing data points to avoid the latest point about artificially implementing values. However, this resulted in limited data and less diversity in county's (high/low economy countries). Even though, there is some variations, most countries with less missing data were countries with economic prosperity.

5.2. Direction of further research

In this study, we tried to predict if GDP goes up or down based on EDI's by using historical data. A direction for further research can be done by examining if EDI's improve GDP level predictions based on numerical time-series data. By comparing the outcomes of an ARIMA time-series model without additional information with a more extended ARIMA model (called Dynamic Regression) we can test if by adding these EDI's the predictions of GDP improve.

The contribution of this study is similar to its academic and societal relevance (see section 1.3 and 1.4): to help aide governing bodies to budget education in a more efficient manner and through this helping their economy blossom, and to help fill certain gaps in the field of data science.

References

Aksoy, S. & Haralick, R.M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters, 22*(5), 563-582.

Appiah, E. (2017). The Effect of Education Expenditure on Per Capita GDP in Developing Countries. *International Journal of Economics and Finance, 9*(10), 136-144.

Assembly, U. G. (1948). Universal declaration of human rights. *UN General Assembly*.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The quarterly journal of economics, 106*(2), 407-443.

Barro, R., & Lee, J. (1994). Sources of economic growth. *Carnegie Rochester Conference Series on Public Policy, 40*, 1-1.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications, 42*(20), 7046-7056.

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance, 47*, 552-567.

Becker, G.S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy, 70*(5), 9–49.

Bellinger, C., Sharma, S., & Japkowicz. (2012, December). *In One-class versus binary classification: Which and when?* Paper presented at the IEEE 11th international conference on machine learning and applications. doi:10.1109/ICMLA.2012.212

Bils, M., & Klenow, P. (2000). “Does Schooling Cause Growth?”. *American Economic Review, 90*(5), pp. 1160-1183.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. Retrieved from https://link.springer.com/article/10.1023/A:1010933404324

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. *In Data mining and knowledge discovery handbook*, 875-886. Springer, Boston, MA.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chin, D., Geweke, J., & Miller, P. (2000) Predicting turning points. Federal Reserve Bank of Minneapolis, Research Department Staff Report 267

Delurgio, S. (1998). *Forecasting: Principles and Applications*. New York, NY: McGraw-Hill.

Ermişoğlu, E., Akçelik, Y., & Oduncu, A. (2013). Nowcasting GDP growth with credit data: Evidence from an emerging market economy. *Borsa Istanbul Review, 13*(4), 93-98.

Feng L., & Zhang J. (2014): Application of artificial neural networks in tendency forecasting of economic growth. *Economic Modelling, 40*, 76-80.

Fitzsimons P. (2017) Human Capital Theory and Education. In: Peters M.A. (eds) Encyclopedia of Educational Philosophy and Theory. Springer, Singapore

Gradstein, M., & Justman, M. (2002). Education, social cohesion, and economic growth. *American Economic Review*, *92*(4), 1192-1204.

Boser, B., Guyon, I., & Vapnik, V. (1992, July). *A training algorithm for optimal margin classifiers*. In Proceedings of the fifth annual workshop on Computational learning theory, New York, NY, USA, 144-152. doi:http://dx.doi.org/10.1145/130385.130401

Hanouz, M. D., & Khatib, S. (2010). The Arab world competitiveness review 2010. Geneva: World Economic Forum.

Hanushek, E. (2016). Will more higher education improve economic growth?. *Oxford Review of Economic Policy*, *32*(4), 538-552.

Hanushek, E., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*(4), 267-322.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software, 45*(7), 1-47.

Honaker, J., G. King, and M. Blackwell. (2018). “Amelia II: A Program for Missing Data.” Retrieved from http://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

Imandoust, S., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications, 3*(5), 605-610.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Retrieved from http://www-bcf.usc.edu/~gareth/ISL/index.html

Junoh, M. Z. H. M. (2004). Predicting GDP growth in Malaysia using knowledge-based economy indicators: a comparison between neural network and econometric approaches. *Sunway Academic Journal, 1*, 39-50.

Kampelmann, S., Rycx, F., Saks, Y., & Tojerow, I. (2018). Does education raise productivity and wages equally? The moderating role of age and gender. *Iza Journal of Labor Economics, 7*(1), 1-37.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of monetary economics, 22*(1), 3-42.

Lutz, W., Creso Cuaresma, J., & Sanderson, W. (2008). The demography of educational attainment and economic growth. *Science, 319*(5866), 1047-1048.

Mandela, N. (2003, July 16). Address by Nelson Mandela at launch of Mindset Network, Johannesburg. Retrieved from http://www.mandela.gov.za/mandela_speeches/2003/030716_mindset.htm

Mankiw, N., & Taylor, M. (2006). *Economics*. London: Thomson.

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. *The quarterly journal of economics, 107*(2), 407-437.

Marković, D., Petković, D., Nikolić, V., Milovančević, M., & Petković, B. (2017). Soft computing prediction of economic growth based in science and technology factors. *Statistical Mechanics and Its Applications, 465*, 217-220.

Montgomery, D., Jennings, C., & Kulahci, M. (2016). *Introduction to time series analysis and* forecasting. Hoboken, NJ: John Wiley & Sons.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *Isprs Journal of Photogrammetry and Remote Sensing, 66*(3), 247-260.

Nelson, R. & Phelps, E. (1966). Investment in Humans, Technological Diffusion, and Economic Growth. *The American Economic Review, 56*, 69-75.

Nowak, A. Z., & Dahal, G. (2016). the contribution of education to economic growth: Evidence from Nepal. *International Journal of Economic Sciences, 5*(2), 22-41.

OECD (2001). Glossary of statistical terms. Retrieved from http://stats.oecd.org/glossary/detail.asp?ID=1264

Ozturk, I. (2001). The Role of Education in Economic Development: A Theoretical Perspective. *Journal of Rural Development and Administration 33*(1), 39–47.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications, 42*(1), 259-268.

Pelinescu, E. (2015). The impact of human capital on economic growth. *Procedia Economics and Finance, 22*, 184-190.

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature. The World Bank. Retrieved from http://datatopics.worldbank.org/education/files/GlobalAchievement/ReturnsInteractive.pdf

Refaeilzadeh, Tang & Liu. (2009). Cross-validation. Retrieved from http://leitang.net/papers/ency-cross-validation.pdf

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2018, September). Nowcasting New Zealand GDP using machine learning algorithms. CAMA Working Paper No. 47/2018. doi: http://dx.doi.org/10.2139/ssrn.3256578

Rieckmann, M., Mindt, L., & Gardiner, S. (2017). Education for Sustainable Development Goals: Learning Objectives. Paris: UNESCO.

Romer, P. (1986). Increasing Returns and Long-run Growth. *Journal of Political Economy, 94*(5), 1002–1037.

Romer, P. (1990). Endogenous technological change. *Journal of Political Economy, 98*(5), 102.

Schultz, T. (1961). Investment in human capital. *The American economic review, 51*(1), 1-17.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003

UNESCO. (2018). *One in Five Children, Adolescents and Youth is Out of School*. Retrieved from http://uis.unesco.org/sites/default/files/documents/fs48-one-five-children-adolescents-youth-out-school-2018-en.pdf

UNESCO. (2014). *Shaping the Future We Want - UN Decade of Education for Sustainable Development (Final report)*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000230171

UNESCO. (2012). *International Standard Classification of Education 2011*. Retrieved from http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

United Nations (n.d.). *Millennium Development Goals and beyond 2015 fact sheet*. Retrieved from https://www.un.org/millenniumgoals/pdf/Goal_2_fs.pdf

United Nations Development Program. (2018). *Human Development Data (1990-2017)*. Retrieved from http://hdr.undp.org/en/data

Vasconcellos, E. (1997). Rural transport and access to education in developing countries: Policy issues. *Journal of Transport Geography, 5*(2), 127-136.

Wals, A., & Benavot, A. (2017). Can we meet the sustainability challenges? The role of education and lifelong learning*. European Journal of Education, 52*(4), 404-413.

Witte, K. D., & López-Torres, L. (2017). Efficiency in education: a review of literature and a way forward. *Journal of the Operational Research Society, 68*(4), 339-363.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation, 1*(1), 67-82.

World Bank. (2018*). Education Statistics Query*. Retrieved from http://datatopics.worldbank.org/education/

Appendix

Appendix I: Evaluation metrics for classification

True Positives (TP)	instances in the positive class, identified as positive
True Negatives (TN)	instances in the negative class, identified as negative
False Positives (FP)	instances in the negative class, identified as positive
False Negatives (FN)	instances in the positive class, identified as negative
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision (PR)	TP / (TP + FP)
Recall (RE)	TP / (TP + FN)
F1	2 x (PR * RE) / (PR + RE)