

Group Project  
Documentation: part 3

Authors: *Agata Makarewicz, Jacek Wiśniewski*

Thesis title: *Application for Analysis of the Economic Growth   
Indexes for European Countries*  
Supervisor: *Agnieszka Jastrzębska, Ph.D. Eng.*

version 1.0

20.11.2021

Table of Contents

[1 Abstract 3](#_Toc87292433)

[1.1 History of changes 3](#_Toc87292434)

[2 Vocabulary 4](#_Toc87292435)

[3 Solution Proposal 5](#_Toc87292436)

[4 Data understanding and preparation 5](#_Toc87292437)

[4.1 Collect initial data 5](#_Toc87292438)

[4.2 Describe data 7](#_Toc87292439)

[4.3 Explore data 7](#_Toc87292440)

[4.4 Data quality 8](#_Toc87292441)

[4.5 Construct data 9](#_Toc87292442)

[4.6 Integrate data 9](#_Toc87292443)

[4.7 Select data 9](#_Toc87292444)

[4.8 Preprocess data 10](#_Toc87292445)

[5 GUI Design 12](#_Toc87292446)

[6 Technology selection 14](#_Toc87292447)

[7 References 14](#_Toc87292448)

[8 Bibliography 14](#_Toc87292449)

# Abstract

This document contains model descriptions for the engineering group diploma thesis entitled “Application for Analysis of the Economic Growth Indexes for European Countries”. It consists of the following parts:

* Solution proposal – short description of the project flow
* Data understanding and preprocessing – detailed information about first steps in the project
* GUI design - user interface vision
* Technology selection - languages, libraries, platforms and other technologies used

As the continuation of the previous document „Group Project Documentation: part 2”, mentioned chapters provide more detailed information about the realisation of the project. The aim is to give guidance to the potential developer so that it is possible to reconstruct the process from scratch. Furthermore, there is a whole chapter dedicated to data understanding and preprocessing which are the steps in the Cross-industry standard process for data mining (CRISP-DM) method.

## History of changes

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Author** | **Description** | **Version** |
| 14.11.2021 | Agata Makarewicz | Template | 1.0 |
| 20.11.2021 | Agata Makarewicz,  Jacek Wiśniewski |  | 1.1 |
|  |  |  | 1.2 |
|  |  |  | 1.3 |

Obraz zawierający tekst

Opis wygenerowany automatycznieObraz zawierający tekst

Opis wygenerowany automatycznie

# Vocabulary

**Homepage** - a webpage presented after turning on the application. It will have all of the functionalities like filtering data and generating the report.

**“Read about the project" page** – a webpage that will present all of the information about the project, authors and contact email addresses.

**Report –** content from homepage consisting of charts and results of clustering algorithms with comments.

**Clustering** - the task of dividing a set of objects into several groups called clusters in such a way that objects within the same cluster are more similar to each other than to objects in other clusters.

**Business cycle** - intervals of expansion followed by a recession in economic activity. Fluctuations are usually characterized by general upswings and downturns in a span of macroeconomic variables.

**Segmentation** – i.e. **time-series segmentation** is a method of [time-series analysis](https://en.wikipedia.org/wiki/Time_series#Analysis) in which an input time-series is divided into sub-series (sequences) with hypothetically homogeneous statistical properties.

**Model** – machine learning algorithm used for clustering.

**Cross-industry standard process for data mining (CRISP-DM) –** process model with six phases describing standard data mining methodology

**K Nearest Neighbors (KNN) –** imputation algorithm. K chosen neighbours’ values are used to calculate new estimates to be imputed.

**Mean squared error –** error measure. It is calculated by the equation

**Interpolation –** imputation algorithm. It has different methods specifying how the data should be imputed. For instance, the linear method fills the gap between known data linearly.

**Correlation** - statistical relationship between two random variables. In this project it is calculated using Pearson’s coefficient, defined as follows: , where X,Y represent variables, σ - standard deviation, and *cov* stands for covariance

**Correlation matrix** – table containing correlation for pairs of variables

**UNDP** – United Nations Development Programme, a United Nations organization which is helping countries eliminate poverty, exclusion and inequalities

**HDI** – Human Development Index, a summary measure of health, education, and economic conditions, developed by UNDP

**PWT** – Penn World Table, a dataset containing many important economic indicators, developed by researchers from the University of Groningen and University of California

**ISO Code** – 3-letter (alpha-3) country code defined by ISO 3166-1 standard

# Model descriptions

In this chapter there are explained three clustering algorithms that are to be used in the final application.

DTW -

## K-Means

One of the algorithms used in the application to cluster countries is K-means. Its purpose is to partition n countries into k, previously chosen, clusters selecting the closest cluster center. Distance between countries was calculated using Dynamic Time Warping (DTW) metric

## Agglomerative clustering

Firstly, each observation (country) is assigned to its own cluster   
Recursively merges pair of clusters

Parameters:

* *n\_clusters* – number of clusters to form
* *linkage* – type of linkage criterion, i.e. the approach to be used for computing the distance between two clusters; the pairs of clusters that minimize this criterion (are closest to each other due to chosen criterion) are merged
  + *“average”* – uses the average of the distances of each observation of the two sets.
  + *“complete”* – maximum value of all pairwise distances between the elements in first cluster and the elements in second cluster
  + *“single”* – minimum value of all pairwise distances between the elements in first cluster and the elements in second cluster
  + *“ward”* – minimizes the variance of the clusters being merged.

## DBSCAN

* Some countries do not have data for the last or the first measured date.

# References

[1] Kaushik, S. (2016). *Analytics Vidhya*. An Introduction to Clustering and different methods of clustering: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

[2] Robert C. Feenstra, Robert Inklaar, Marcel P. Timmer. (2021). *Penn World Table*. Pobrano z lokalizacji https://www.rug.nl/ggdc/productivity/pwt/

[3] *United Nations Development Programme*. https://www.undp.org/

[4] *Wikipedia*. Business cycle: https://en.wikipedia.org/wiki/Business\_cycle

# Bibliography

1. Aghabozorgi, Saeed, Shirkhorshidi, Ali S., and Wah, Teh Y. Time-series clustering – A decade review. Information Systems 53 16-38, 2015.
2. Gräbner, C., Heimberger, P., Kapeller, J., and Schütz B. Structural change in times of increasing openness: assessing path dependency in European economic integration. Journal of Evolutionary Economics 30, 1467–1495, 2020.
3. Bartlett, W. and Prica, I. Interdependence between Core and Peripheries of the European Economy: Secular Stagnation and Growth in the Western Balkans. LSE‘Europe in Question’ Discussion Paper Series, LEQS Paper No. 104/2016, 2016.
4. Hamilton, James Douglas Time Series Analysis. Princeton University Press, 1994.
5. Pal, Avishek, Prakash, PKS. Practical Time Series Analysis. Master Time Series Data Processing Visualization and Modelling Using Python. Packt, 2017
6. Fu-Lai Chung, Tak-Chung Fu, V. Ng and R. W. P. Luk, "An evolutionary approach to pattern-based time series segmentation," in IEEE Transactions on Evolutionary Computation, vol. 8, no. 5, pp. 471-489, Oct. 2004.