Jean Monnet University

Machine Learning and Data Mining master program

Second Semester

“Knowledge Discovery and Data

Mining”

Data Mining Project with R

*Author*

Kristina Kulivnyk

*Supervisor*

Prof. Fabrice Muhlenbach

March 13, 2019

Table of Contents

[Problem understanding 3](#_Toc3382936)

[Data understanding 4](#_Toc3382937)

[Data preparation 5](#_Toc3382938)

[Modeling 6](#_Toc3382939)

[Evaluation 7](#_Toc3382940)

[Deployment 8](#_Toc3382941)

[Sources: 9](#_Toc3382942)

# Problem understanding

For Data Mining project I decided to choose a dataset that combine information about state of education in Ukraine. It has diverse data, starting from … up to … .

I’ve choose this dataset because it is compelling and indispensable personally to me to get a deeper understanding of the state of education in my native country via building various statistical and predictive models. The reason of such interest is in an increasing number of debates over the quality of education in Ukraine, its relevance comparing to European systems and causes that could help to bring it to a new, more advanced level. Since posting a real state or the education has lots of rigid drawbacks for the national government, there exist a lot of populism and manipulation both in press and web articles over such vulnerable topic. My goal is to get relevant information and to build a genuine models using independent data, collected by UNESCO.

I hope that my work could not only provide some valuable information personally for me but would also be a good start for actual changes in educational system in Ukraine. I believe that when we know reasons, we could fight for better results.

The data for my research was found on a UNESCO website (could be accessed via the link - <https://data.humdata.org/dataset/unesco-indicators-for-ukraine> ). The goal of such datasets is to provide the most recent information about state of culture, education, literacy of the country. Datasets are updated every year.

# Data understanding

Taking the first glance over my data I decided to check several questions :

1. Discover patterns about preferred fields of studies for a male/female student over the years (2012-2018) and make a prognosis about 2019.
2. Make cluster analysis of the countries where Ukrainian student will preferably go depending on educational level (bac, license, master) over the years 2012-2018. Make prognosis about 2019
3. Make an analysis of teachers’ qualification over the years, depending on sex and educational institutions level.
4. Discover attendance patterns for urban and rural locations.

# Data preparation

# Modeling

# Evaluation

# Deployment

To make the work with the project more structured and efficient I’ve created a github repository that hosts report file, data and R code itself. It could be accessed via the link: <https://github.com/KrisKuliv/MLDM_Data_Mining>

There exist three branches:

* **Master branch** with the final version of the project
* **Docs branch** where only report file is stored
* **Data\_code branch** where R project files are stored

I’ve also created an R package that can be accessed via the command:

Working with git repository helped me to keep all changes secure and not to be afraid of loosing any precious results. It is also interesting to check the statistics of commits to understand the personal productivity level.

As using git is a best-practice in the biggest part of IT companies, working with this project also made me train my git-usage skills.

# Sources:

1. Ukraine - Sustainable development, Education, Demographic and Socioeconomic Indicators - Humanitarian Data Exchange. *Data.humdata.org*, 2019. <https://data.humdata.org/dataset/unesco-indicators-for-ukraine>.
2. Google's R Style Guide. *Google.github.io*, 2019. <https://google.github.io/styleguide/Rguide.xml#filenames>.
3. R, R. Consistent naming conventions in R. *R-bloggers*, 2019. <https://www.r-bloggers.com/consistent-naming-conventions-in-r/>.
4. Romeo, V. Preparing the data for modelling with R. *R-bloggers*, 2019. <https://www.r-bloggers.com/preparing-the-data-for-modelling-with-r/>.
5. 11 Most Useful Steps to Create Data Exploration in R | Methods | Example | Definition. *EDUCBA*, 2019. <https://www.educba.com/data-exploration-in-r/>.
6. Data Preparation — A crucial step in Data Mining. *Medium*, 2019. <https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281>.
7. The 10 Mining Techniques Data Scientists Need for Their Toolbox. *Towards Data Science*, 2019. <https://towardsdatascience.com/the-10-mining-techniques-data-scientists-need-for-their-toolbox-ae15a5733b02>.
8. 8 Data Mining Techniques You Must Learn To Succeed In Business. *Medium*, 2019. <https://medium.com/@onix_systems/8-data-mining-techniques-you-must-learn-to-succeed-in-business-ae4032bf6469>.
9. 15 Easy Solutions To Your Data Frame Problems In R. *R-bloggers*, 2019. <https://www.r-bloggers.com/15-easy-solutions-to-your-data-frame-problems-in-r/>.
10. r, W. Clustering Mixed Data Types in R. *R-bloggers*, 2019. <https://www.r-bloggers.com/clustering-mixed-data-types-in-r/>.
11. blog, A. 5 ways to measure running time of R code. *R-bloggers*, 2019. <https://www.r-bloggers.com/5-ways-to-measure-running-time-of-r-code/>.
12. Factoextra R Package: Easy Multivariate Data Analyses and Elegant Visualization - Easy Guides - Wiki - STHDA. *Sthda.com*, 2019. <http://www.sthda.com/english/wiki/factoextra-r-package-easy-multivariate-data-analyses-and-elegant-visualization>.
13. K-Means Clustering in R: Algorithm and Practical Examples - Datanovia. *Datanovia*, 2019. <https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/>.
14. K-means Clustering in R with Example. *Guru99.com*, 2019. https://www.guru99.com/r-k-means-clustering.html.
15. Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA. *Sthda.com*, 2019. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r.
16. Linear Regression With R. *R-statistics.co*, 2019. http://r-statistics.co/Linear-Regression.html.
17. Quick-R: Multiple Regression. *Statmethods.net*, 2019. https://www.statmethods.net/stats/regression.html.
18. 5 Amazing Types of Clustering Methods You Should Know - Datanovia. *Datanovia*, 2019. https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/.
19. Hierarchical Cluster Analysis · UC Business Analytics R Programming Guide. *Uc-r.github.io*, 2019. https://uc-r.github.io/hc\_clustering.
20. Cluster Analysis in R. *Xperimentallearning.blogspot.com*, 2019. http://xperimentallearning.blogspot.com/2018/04/cluster-analysis-in-r.html.
21. Performing a cluster analysis in R. *Instant R*, 2019. http://www.instantr.com/2013/02/12/performing-a-cluster-analysis-in-r/.
22. (first.last@ucr.edu), F. Cluster Analysis in R. *Girke.bioinformatics.ucr.edu*, 2019. http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html#2\_data\_preprocessing.
23. RPubs - Cluster Analysis in R: Examples and Case Studies. *Rpubs.com*, 2019. https://rpubs.com/gabrielmartos/ClusterAnalysis.
24. Josse, J. Handling missing values with R. *Juliejosse.com*, 2019. http://juliejosse.com/wp-content/uploads/2018/06/DataAnalysisMissingR.html#4)\_multilevel\_(mixed)\_data\_with\_missing\_values.