COMP 5411 FA / FB
Fall 2019
Project Proposal

Project Title: Erasmus Mobility Statistics

Group Members:

Name Student	Student No.	Section
Anna Dieckvoss	1127324	FB
Borys Komarov	0889421	FB

Big Data Project Proposal

Anna Dieckvoss (1127324) and Borys Komarov (0889421)

Abstract— **This is the project proposal for the project of the lecture Topics in Big Data at Lakehead University in Fall 2019. The course is part of the graduated studies in Computer Science (5411-FB) and is given by the Instructor Dr. Quazi Abidur Rahman.**

I. INTRODUCTION

A. Problem definition

The topic of this paper is the Human Mobility Pattern Clustering. Based on an Erasmus data sets we analyzed the student & staff mobility trends in 2012-2013.

Erasmus exchange program operates since the late 1980 's and have sent 3.000.000 students to another country. It stands for EuRopean Community Action Scheme for the Mobility of University Students. With Erasmus students get the opportunity to study at universities in the EU member states for set periods of time. The generated credit points can be transferred to the home university [1]. The Bologna reform defines a standardization of study in the European Union. The Erasmus program is also extended for staff working in education non-necessarily on teaching positions. Participants of the project describe that it was very hard to organize the process although there is a general standardization and it was definitely worth going for it and changed their life for good.

The mobility of young people is interesting to know because based on analyzing human mobility patterns we can predict destination choices for the following year and improve the overall experience for students and staff on the Erasmus program by providing them with a recommendation on the most suitable destinations countries. More and more people want to try working or gaining knowledge abroad every year in our study we try to help students and staff enrolled in the Erasmus program to make more informed decisions.

B. Significance of the problem

The number of students and staff who are going abroad every year is highly increasing. The broad social, economic and political factors of this mobility are diverse and not always known. It was very nicely put in [5] “Experiences and meanings gained in a mobility program like Erasmus have implications beyond academic achievements.”. Globalization impacts the world and more and more companies become worldwide having offices in different locations with different cultures and one should be prepared for it after graduation. Academic mobility is an opportunity for students and staff to gain new knowledge, experience different cultures, and build new connections.

The more people go on the Erasmus and similar programs the more data is generated and it now becomes even more important to provide accurate and automated recommendations. The recommendation will be based on initial clustering on the patterns and then analyzing them together with the various standard features including destination country, origin country, study program, program length, credit points. Potentially, our findings could help hundreds of thousands of exchange or potential-exchange students in Europe. Many parties can benefit from our study. Countries and universities might get a detailed analysis of how they are performing in terms of international student recruitment and what would be predictions for them for the nearest future. Students can get a chance to maximize their experience by adopting the power of data on their side. Erasmus program may get deeper insights on which countries are better encouraging incoming traffic of international students and staff and what affects that.

C. Literature review

We were able to find several articles exploring various aspects of the Erasmus program and attempting to produce insights. The paper [3] analyze the mobility network of Erasmus staff and students within a big European project FETCH (Future Education and Training in Computing). The goal was to evaluate the current state of the Erasmus mobility network among institutions. The authors applied Techniques that are commonly used in social network analysis to academic mobility flow. “The structure of the network is investigated using connected component analysis and k-core decomposition.”

At the same time authors in another paper [4], which uses the same data as we are using for our study, but for a different time period. They were trying to cluster countries based on the number of students coming into and out of the country and then make a conclusion if there is a finite number of classes as well as whether every country falls into this set of identified classes. In their results, they present that it is possible to represent all countries using three classes: good importers and exporters, good importers only and good exporters only. Their results are quite interesting, however, they were not taking into account more than a couple of features from the entire dataset which gave them results “consistent with previous studies”[4], but almost no interesting insights or breakthroughs.

D. How the proposed project matches/differs with previous work

There are several things in common between the research we found and our proposal as most of the research we found is based on custom collected data or the open datasets similar to the ones we are going to use, however, they either take only one aspect of the data (for example only students mobility) [4] whereas our proposal is to go above just looking at the problem from the student's perspective and that is why we decided to extend our study for the staff placements as well. The topic itself is not extremely popular and therefore the research on the topic is sparse and spread along the last 20 years and as a result, there is no actual research on the up-to-date data. Some papers focus on a broader range of countries, not only within the European Union [3]. In this paper, we study the Erasmus student mobility and not especially computer science programs focus like the paper [3].

During our investigation of the existing literature we often faced two extremums where most of the papers are presenting qualitative, descriptive research, but, on the other hand, others which we found were focused only on computer science aspects of the data [3]. That is why we aim to make our research using both descriptive analytics and computer science methods so it could fill in the gap between descriptive and purely technical research on the topic.

II. DATA

A. Data source

The used datasets are the official Erasmus mobility statistics 2012-13, which has been published in 2015 on the open dataportal of the EU [2]. It is delivered as a csv file.

B. Data description

Dataset representing students mobility includes 34 Features such as the student ID, home institution and ECTS Credits and 248153 Instances represent the exchange students. It has both categorical (student subject area, gender) and continuous (length of the study period, number of ECTS credits) features. The file contains missing values, which are marked witch '?' or Unknown.

If we have time to work on the comparison between student, staff, and teachers mobility we are going to also use two additional datasets from the same source one for staff mobility and another one for teachers mobility. Staff mobility dataset has 16556 Instances and 23 Features whereas teachers mobility dataset has 36071 Instances and 25 Attributes. Both of them contain some amount of missing values.

III. METHODS & TOOLS (TENTATIVE)

A. Anticipated preprocessing steps

Our dataset includes some missing values, so we would like to implement a couple of imputation methods such as k-NN or wk-NN to compare the performances and choose the best. Additionally, as it is very common for our features to come from different domains we will need to normalize them in order to apply various methods on multiple features. What also may occur during the normalisation process is that we would like to add more weight to some features and that should be anticipated as well. Having 34 features on the primary dataset we would have to do some reductions to the number of features in use.

B. Anticipated learning methods

Considering the type of data we are working on and the nature of the topic we are definitely going to use some descriptive statistics methods to describe our data and produce insights. Both raw data and insights from the descriptive statistics can then be used when applying more advanced automated methods. We currently have two methods under review Clustering and Regressing. Our initial idea would be to use clustering to produce initial insights and regression as a tool for predictions. During our research we are going to use multiple clustering algorithms such as k-means clustering, Agglomerative Hierarchical Clustering, etc... and then compare and report the results for different methods.

C. Anticipated validation methods

As we are going to use data from 2012/2013 we may make an inquiry for the latest data and compare our results for predictive analytics with the actual up-to-date data. Additionally, we are going to look at the existing research and compare our findings with the existing ones.

D. Tools

We decided to use R programming language and R-studio IDEA. R comes with a good number of methods embedded within its standard library, however, we would still need to make an investigation on the existing popular data science libraries for R. One definite candidate we are going to use is data.table library as our dataset is quite big and the library is more efficient in working with larger datasets than standard data.frame included in R.

REFERENCES

[1] https://www.erasmusprogramme.com/post/what-is-the-erasmus-programme

[2] https://data.europa.eu/euodp/en/data/dataset/erasmus-mobility-statistics-2012-13

[3] Analysis of ERASMUS Staff and Student Mobility Network within a Big European Project, Miloš Savić, Mirjana Ivanović, Zoran Putnik, Kemal Tütüncü, Zoran Budimac, Stoyanka Smrikarova, Angel Smrikarov,MIPRO 2017, May 22- 26, 2017, Opatija, Croatia

[4] Breznik, Kristijan & Skrbinjek, Vesna & Law, Kris & Đaković, Goran. (2013). ON THE ERASMUS STUDENT MOBILITY FOR STUDIES.

[5] Mizikaci, Fatma; Arslan, Zülal Uğur. A European Perspective in Academic Mobility: A Case of Erasmus Program. Journal of International Students . 2019, Vol. 9 Issue 2, p705-725. 21p.