COMP 5411 FA / FB

Fall 2019

Final Project Report

Project Title: Student Mobility Pattern in the EU

Group Members:

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

Student Mobility Pattern in the EU

Anna Dieckvoss (1227324)
Department of Computer Science
Lakehead University
adieckvo@lakeheadu.ca

Borys Komarov (0889421)
Department of Computer Science
Lakehead University
bkomarov@lakeheadu.ca

ABSTRACT

Background: The analysis of human mobility patterns helps to predict the destination choices. Prediction of the most suitable destinations countries improves the overall experience for students in the Erasmus program.
Objective: This study aimed to use Descriptive analytic and machine learning methods to (1) predict suitable destinations countries for future students and (2) measure relative performance of Universities to help the Erasmus coordinator to make informed finical decisions.
Methods: The destination choices are defined as the 33 countries, which are participating as host countries in the Erasmus program. Prediction models were developed using 2 methods: (1) Naive Bayes (2) Random Forests. ... The relative performance of Universities defined by comparing different success steps and grouping with the method k medoids.
Results: In this class-imbalanced dataset, all 2 clustering models achieved (1) top-1 accuracy: 58.587246%, top-3 accuracy: 74.547705% and (2) top-1 accuracy: 58.677111%, top-3 accuracy: 66.642929%. The prediction model based on Random Forests performs the best
Conclusions: Classification analysis was applied to predict the top 3 destination for students in the EU. using machine learning methods based on. Clustering analysis was applied to group Universities based on their relative performance.

Keywords: Erasmus Dataset, classification analysis, cluster analysis, Student Mobility Pattern

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 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 on the Erasmus program by providing them with a recommendation on the most suitable destinations countries. More and more people want to try gaining knowledge abroad every year in our study we try to help students enrolled in the Erasmus program to make more informed decisions.

B. Objectives

This study aimed to use Descriptive analytic and machine learning methods to predict suitable destinations countries for future students by using classification. Additionally, the relative performance of Universities was measured by comparing different success steps to help the Erasmus coordinator to make informed finical decisions based on the clustered university.

C. Significance of the problem

The number of students 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 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 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 what affects that.

D. 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.

E. How the proposed project matches 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 data sets 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 to propose a view from ERASMUS organisation perspective. 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.

The dataset students mobility dataset contains 3 types of mobility:

- Study mobility
- Placement mobility
- Study + Placement mobility

Not all of the attributes are present for all 3 mobility types. The distribution of instances are following. Study mobility - 211,519 instances, Placement mobility - 55,552 instances, Study with placement mobility - 476. During this project we primarily focused on Study mobility type combined on the study part with the study with placement mobility type.

III. Methods & Tools

A. Initial data prepossessing

While studying our dataset we realised that there is a need for clean up of the data. As we decided to focus on study placements we first had to filter those from the whole dataset. Then some adjustments had to be made to remove unnecessary and meaningless attributes. The following actions were taken

1) Garbage attributes reduction: Most of the attributes in our dataset do not have self-explanatory names, therefore there is an additional dictionary supplied along with the dataset. While studying the dictionary we realised that the following attributes are unnecessary and do not play important role for further analysis.

- ID_MOBILITY_CDE - ID of a placement
- CONSORTIUM_AGREEMENT_NUMBER - Abstract agreement number not relevant for our objectives
- SPECIAL_NEEDS_SUPPLEMENT_VALUE - As we do not separate students with / without special this attribute is not useful for us
- SHORT_DURATION_CDE - This attribute shows whether student returned home due to "force majeure". We do not consider this special case

- QUALIFICATION_AT_HOST_CDE - Whether student will receive any kind of qualification (degree) at the host university. This attribute is not relevant to our objectives as it depends not on a university, but mostly on time student is participating in the program.

2) Work placements attributes reduction: As we consider only students with study placements we need to get rid of all attributes related to work placement information.

- MOBILITY_TYPE_CDE - Flag used to separate study and work placements. We no longer need it because we have already selected only study placements.
- PLACEMENT_ENTERPRISE_VALUE
- PLACEMENT_ENTERPRISE_CTRY_CDE
- PLACEMENT_ENTERPRISE_SIZE_CDE
- TYPE_PLACEMENT_SECTOR_VALUE
- LENGTH_PLACEMENT_VALUE
- PLACEMENT_START_DATE
- ECTS_CREDITS_PLACEMENT_AMT
- PLACEMENT_GRANT_AMT

3) Missing values in our dataset: Missing values in our dataset are represented as 0 if attribute is continuous and as one of the following strings for categorical attributes.

- "? Unknown ?"
- "???"
- "?"

As most of the remaining attributes in our dataset are categorical we manually checked all the continuous ones and wrote a procedure to programmatically check all the categorical ones. If any missing values were detected the warning would be printed to the console. We were not able to find any missing values in the selected subset of attributes. Most missing values in the dataset are caused by the fact that it contains 3 types of placement instances and there are some attributes which are not shared among all the 3 instance types.

4) Attributes clean up: Finally, we noticed a few attributes which needed some cleaning. First of all, there was one country in the dataset represented by three different country codes and so we grouped all three of them under one code. It was Belgium which was initially split into:

- "BEDE"
- "BEFR"
- "BENL"

We grouped them under "BE" code. Secondly, many languages were represented by multiple spellings of the language code, for example initially we could have the following options for English language "En", "EN", "en". This was fixed by making all language code values upper case. Lastly, the "STUDY_GRANT_AMT" attribute, which represents total amount of scholarship support received by student, was initially stored as character string representing an integer number. We had to convert it to numeric field taking into consideration two possible decimal positions.

B. Classification

One of the primary goals we tried to achieve was to improve overall user experience with the ERASMUS program and especially first steps when student needs to select destination country and university by trying to suggest best options based on the similar student profiles from the previous year. This hypothetical program can be used both by students themselves or by ERASMUS program coordinators at each university participating in the program. Our task was to predict "HOST_INSTITUTION_COUNTRY_CDE" attribute values.

1) Preprocessing: For classification task we did several things at data preprocessing stage such as removing attributes directly correlated with the target variable, features selection and trying to address class imbalance. We found out that "HOST_INSTITUTION_CDE" is directly correlated with the target variable and therefore we had to remove it from the list of used features.

2) Features selection: For both classification methods described below we used the same features selection algorithm: Sequential Forward Selection[6] (SFS). SFS is a simple features selection method which starts with an empty set of features at each iteration it tries to find the best feature and add it to the set. We repeat process until there is not performance gain. To validate features selection method performance we used Objective Function[6] with wrapper approach using Naive Bayes as validation Machine Learning method.

3) Addressing class imbalance: During the preprocessing stage we decided to investigate our classes in details. The first thing we could see was that there exist a significant class imbalance.

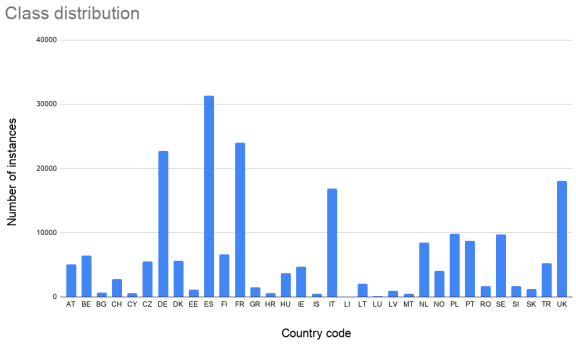

Fig. 1.

4) Naive Bayes: The first classification method we applied to our problem was Naive Bayes. Naive Bayes is a supervised machine learning algorithm which uses Naive version of Bayes Theorem [7]. I counts conditional probability for each possible value of the target variable and then selects the value with highest probability as a result. We decided to use it, because it works well with categorical data and is easy to understand and implement.

5) Random Forest: The second classification method we selected was Random Forest [9]. Random Forest combines Begging and Decision Tree into single supervised Machine Learning method. It is achieved my building some amount

	Accuracy [%]			F1 score
	top-1	**top-2**	**top-3**	**top-1**
Naive Bayes	58.58	68.90	74.54	0.33
Random Forests	58.68	63.84	66.65	0.32

TABLE I
CLASSIFICATION RESULTS

of trees using training dataset and then running test instances down those trees in order to classify them.

6) Validation methods:

C. Clustering

k-means clustering for partitioning a data set into k groups or clusters. In k-medoids clustering, each cluster is represented by one of the data point in the cluster. K-medoid is a robust alternative to k-means clustering. This means that, the algorithm is less sensitive to noise and outliers, compared to k-means, because it uses medoids as cluster centers instead of means (used in k-means). PAM algorithm (Partitioning Around Medoids, (Kaufman and Rousseeuw 1990)).

D. Tools & Libraries

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 still needed to include some libraries.

We use the following libraries in our project:

- naivebayes - Naive Bayes implementation
- randomForest - Random Forest implementation
- ...

E. Features for Prediction Model

Prediction Models Measuring Prediction Performance Class Imbalance

IV. RESULTS

A. Descriptive analytic

...Placement (55552) vs Studies (211519) amount and combined (476) Top sending countries Spain(39249) France(35311) Germany(34891), Italy, Poland Top receiving countries/ institution Spain(31360) France(23970) Germany(22728) UK Italy Average Duration is 6.2 Month for S Thought Languages English (117260) Spanish(23419) French (22437)

B. Classification Results

...
Random Forest top-2 accuracy:
Random Forest top-3 accuracy: 66.642929

C. Clustering Results

...

V. CONCLUSION

...

A. Future Work

In future, we shall focus on the comparison between student, staff, and teachers mobility placements. There is also the opportunity to compare the results of 2012-2013 to other years.

ACKNOWLEDGMENT

We would like to thank Dr. Quazi Abidur Rahman for guiding us through this research study and enlightening us with his in depth knowledge.

APPENDIX I
REPRODUCTION OF OUR WORK

We used R code in Rstudio ... Attached to this document is a zip file with all the Code. After downloading the R Studio version 1.2.1335, you can execute the code by running the main(). The main function has install the following packet naivebayes, randomForest (https://github.com/majkamichal/naivebayes) install.packages(c("cluster", "factoextra")) three primary program modes implemented (value for program_mode parameter): * Classification - 0 * Clustering - 1 * Descriptive analytics - 2

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 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.
[6] http://research.cs.tamu.edu/prism/lectures/pr/pr_l11.pdf
[7] https://plato.stanford.edu/entries/bayes-theorem/
[8] https://github.com/majkamichal/naivebayes
[9] https://www.stat.berkeley.edu/ breiman/RandomForests/cc_home.htm
[10] https://cran.r-project.org/web/packages/randomForest/randomForest.pdf