Google Summer of Code Gender Diversity: An analysis of the last 4 editions

Jhemeson Silva Mota
University of Brasilia (UnB)
Brasilia, Brazil
jhemesonmotta@gmail.com

Edna Dias Canedo
University of Brasilia (UnB)
Brasilia, Brazil
ednacanedo@unb.br

Marcio Vinicius Okimoto
University of Brasilia (UnB)
Brasilia, Brazil
marciobtos@gmail.com

Jhonatan Silva Mota
Lutheran University Center of Palmas
Brasilia, Brazil
savaegt@gmail.com

ABSTRACT

This work presents a comprehensive research about the participation of men and women in the area of Information and Communications Technology (ICT) through data extracted from the last four editions of Google Summer of Code (GSoC). The goal of this work is to find Association Rules between gender characteristics and coding using the Apriori Algorithm. A total of 61 association rules were generated through the aforementioned algorithm, being 22 of them found only in the data set with the women, 24 found only with the men, and 15 applicable to both sets. We can cite as one of the main findings of this work the fact that the representativeness of women in GSoC is decreasing in the last few years. Despite this, the representativeness of women in GSoC is above average, according to what has been reported in other studies in the literature in which women are underrepresented. When it comes to the most utilized technologies, we have “Python", “Java", “C++", “C" and “JavaScript" in the top. Analyzing technologies, it’s possible to realize that the main utilized technologies for men and women are similar, but, in general, men are more likely linked to programming languages. The most common project topics are: “Event Management", “Web", “Web Development", “Data Science" and “Cloud" in the top. This can represent how diverse the project topics of the database are, but not necessarily has something related to gender.

KEYWORDS

Data mining, Gender diversity, GSoC, Apriori, Association rules

1 INTRODUCTION

In the last two decades the Information and Communications Technology (ICT) area is expanding in a fast-track pace. Globalization and new technologies demand a large number of professionals favorably inclined to study and consequently work in this field [41]. Even with many years of struggling for gender equity, in the current scenario there are few women who are enrolling in science, technology, engineering and mathematics (STEM) courses, both undergraduate and graduate level [34]. The related number is especially low in computer related courses, such as Computer Science (CS) and Computer Engineering (CE). In actual fact, this number - women enrolling in STEM related disciplines - is decreasing, regardless of all the efforts. Thus, CS and CE remain fields predominated by male [5]. There has been an increasing number in studies approaching women in a men-dominant field. Women have entered many other fields where previously there were no considerable female representativeness, including other STEM fields, but not computer science and engineering [12]. Amplifying women interest in STEM, specifically CS, is an important aim for universities, national and local governments, and society as a whole [30].

In other study presented by Moudgalya et al. [29] data mining is used to explore Stack Exchange - a Question and Answer (Q&A) forum, specifically the *Computer Science Educators Stack Exchange* (CSEd SE) [17] to understand and analyze the view of computer science educators on gender diversity by using a non-intrusive technique on the forum. In a work presented by Botella et al. [6], they argue that one action driven to reduce the gender gap is the gendered innovation initiative and, in particular, Machine Learning and Data Science areas suppose new opportunities to include gendered innovation in Information Theory. Which could be achieved by the fact that they can be applied to many different domains. A large quantity of research has investigated key factors that influence the interest of female students in STEM and CS (e.g., [10], [46], [6]).

LinkedIn’s 2018 Diversity Annual Report shows that women represent 42.9% of workforce of LinkedIn, with a representativeness of 39.1% in leadership positions, representing a 12% increase of women in leadership positions over the last two reports [16]. The 2019’s report shows that women represented nearly 41% of the company’s leadership (an increase of 17% in the last three years and 56% in the last five); 22% of the technical roles; and 55% of the non-technical roles [13]. Despite this increase, women representation in the technology area is 21.8%, while men represent 78.2%. Some organizations aim to change this scenario of women under-representation, with programs committed to a set of goals to increase the representation of the women workforce and create a more inclusive culture. For example, a recent Google Enterprise Diversity report shows that hiring of women has increased to 33.2% [7], an increase of more than 1.9 points in relation to the previous survey. In particular, the concentration in non-tech areas increased to 47.2%, representing an increase of more than 3.3 points. Although hiring of women increased, hiring in the company for leadership positions decreased by 25.9% (-3.5 points) [7] in relation to the previous report.

This work presents a comprehensive research about the participation of men and women in the area of Information and Communications Technology with focus on the open-source community, using data extracted from the last four editions of GSoC. The purpose of this research is to find association between gender characteristics and project characteristics using Association Rules. With that perspective over the association rules, it will be possible to identify if there is discrepancy between the contextual standards of men and women.

This article is organized as follows. Section 2 a contextualization of the related subjects to this research is presented. Section 3 presents the methodology utilized for the development of the work. After that the results of the statistical analysis and of the data mining that were made are presented on Section 4. The identified possible threats to validity and limitations are presented on Section 5. Lastly, final considerations and suggestions for future work are presented in the Section 6.

2 BACKGROUND

2.1 Google Summer of Code

Google Summer of Code (GSoC) is a global program focused on bringing more student developers into open source software development. Students work with an open source organization on a 3 month programming project during their break from school. Currently the program has more than 14,000 students, 109 Countries, 651 Open Source Organizations and more 35,000,000 Lines of code [21].

Since 2005, a total of 686 open source organizations have been a part of GSoC, bringing new, excited developers into their communities and the world of open source. The program is "open to university students, age 18 and older in most countries" [21]. GSoC is well known among the Summers of Code, and provides its students with a broad range of rewards for participation including: participating in a global company program; community ties; skills development; personal satisfaction; professional advancement; recognition among peers; status; and financial remuneration [39]. In addition to that, the same author says that in GSoC, "the vision and experience of core members of the community influence project selection, and the intensive mentoring process facilitates creation of strong ties". With that, the biggest part of GSoC projects result in stable features.

2.2 Data Mining

Data Mining (DM) is, according to Santos [36], a set of techniques and procedures that have the goal to extract high level information from raw data. In a different definition, Beck et al. [2] says that DM is "the field of discovering novel and potentially useful information from large amounts of data". The work produced by Fayyad et al. [18] describes it as a step contained in the Knowledge Discovery in Databases (KDD) process that consists of performing data analysis and applying discovery algorithms that, under certain computational limitations, produce a set of patterns for certain data.

The aforementioned therm (KDD) refers to the vast process of finding knowledge in data. It's of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The process contains five steps of processing to make this transformation from Data to Knowledge [25]: 1) Selection of data: this step covers the selection of the data set on which the discovery process will be performed; 2) Preprocessing: comprehends the action of removing noise and outliers from the selected data set, as well as the formulation of the strategy to be adopted in case of missing data; 3) Transformation: in this step happens a reduction on the effective number of variables under consideration. Usually are discarded variables with contextual irrelevancy; 4) Data Mining: stands for the definitions mentioned above; and 5) Evaluation: contains the interpretation of mined patterns and, after a good evaluation, the outcome is a consolidated knowledge. The visual flow of the five steps can be seen through the Figure 1.

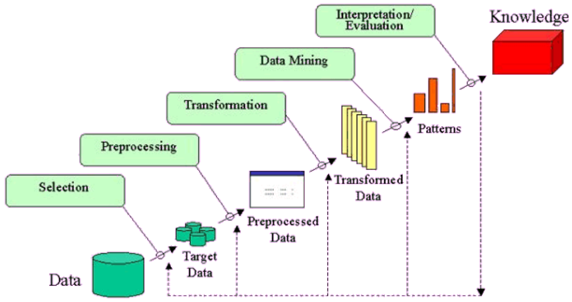

Figure 1: KDD Steps [20]

The goals to be achieved with DM are defined based on the objective of the application, and are classified into two types: a) When you want to verify user-defined hypotheses, the goal is defined as "verification"; b) In cases where patterns are to be set autonomously, the goal is defined as "discovery" - where, in this type the prediction and description tasks are performed [42]. Mining data to discover knowledge isn't a trivial task. You need to know the data, the process of analysis and discovery, the tasks, the data mining techniques, and the mathematical and computational tools that apply in this context [15].

2.2.1 Association Rules. According to Mobasher et al. [27], there is no consensus in the literature regarding the ideal data mining technique for each application, not even the criteria to be used to evaluate different data mining techniques. Among the techniques considered efficient for data mining, there are the association rules, commonly used as a learning technique that doesn't need supervision and can be used to identify novel patterns amongst entities in a large set of data [19]. This technique has the purpose to find links between attributes assuming that the presence of one attribute in an certain event implies the presence of another attribute in the same event [15].

Regarding the concept of this technique, Moonen et al. [28] said that the association rules are "implications of the form A → B, where A is referred to as the antecedent, B as the consequent, and A and B are disjoint sets". Adding to that, Camilo and Silva [8] described the association rules as one of the most utilized and well-known techniques of data mining around the world. Still, according to the same authors, the results of the applying of this technique will provide associative rules between items of the data set. The classic application of the association rules is the "shopping cart", e.g.: if the customer will buy milk and bread, there is a good probability that he will also buy butter.

2.3 Related Works

Some related works can be found at the literature. Babes-Vroman and Nguyen [1] studied the subject of gender diversity within Computer Science at an University with thousands of students. Their results indicate that a large proportion of women who take the Introductory CS1 course for majors do not intend to major in CS, which contributes to a large increase in the gender gap immediately after CS1. The same aspect of gender diversity has been analyzed by Bosu and Sultana [4] in a research that aims to determine the level of gender diversity among popular open source software projects and identify the presence of gender biases that may discourage women participation. This work suggest that the lack of gender diversity remains an issue as each of the ten projects analyzed had less than 10% women developers.

Meanwhile, the work developed by Vachovsky et al. [41] alerts how recent diversity reports demonstrate a large gap between the percentage of women holding computing jobs compared to the percentage of men. In addition, the authors evaluate the Stanford Artificial Intelligence Laboratory's Outreach Summer (SAILORS) as a way of addressing the lack of diversity in Artificial Intelligence. The results show a positive impact of SAILORS by achieving the goals of contextualizing technical AI concepts through social impact and addressing barriers for girls in computer science.

Regarding the impacts of the gender diversity, Blincoe et al. [3] researched with the aim to examine how the working atmosphere depends on the gender diversity of IT teams. Their results appoint that the atmosphere in teams with diversity usually is more pleasant when compared to purely male ones. Moudgalya et al. [29] researched about the perceptions of equity and gender diversity in Computer Science and got results suggesting that "researchers need to continue to examine educator perceptions so that we can design appropriate online teacher communities, teacher education courses, and professional development workshops to address equity and gender diversity issues in CS". Also studying about gender diversity, Hoogendoorn et al. [22] estimate the impact of the share of women in management teams on their business performance. The result of the work states that "management teams with an equal gender mix perform better than male-dominated and female-dominated teams in terms of sales, profits and earnings per share".

Silva et al. [38] analyzed the effectiveness of initiatives such as GSoC and found that 82% of the participants are able to merge successful changes to the desired project, while 40% of students kept contributing longer than a month after the project and 15% contributed longer than a year. Analyzing the same program, Silva et al. [37] also studied what motivates students to enter programs like GSoC and resulted in the discovery that they enter the program to have richer experiences and not necessarily to become frequent contributors on the open-source community.

In this work gender diversity in the field of computer science is studied, however, from the perspective of the areas of statistics and data science - according to which we can look for patterns associated with genders. In addition, as well as in Silva et al. [37], GSoC is used as a research universe - but for different purposes, since the cited work does not have the goal to study gender diversity.

3 METHOD

3.1 GSoC Data Mining

The current project was built utilizing different materials in different steps. As shown in Figure 2, the gender inference process went through a number of steps. First it was collected a set of data from the last four editions of Google Summer of Code web platform through a web scraper - this software was built with the programming language Python utilizing libraries like: Beautiful Soap 4, "a library that makes it easy to scrape information from web pages" [14]; and Pandas [31], an open source data analysis library.

The outcome of the mentioned process was a spreadsheet in CSV format. From this file, in order to make the gender inference of people, we used two different tools together: GenderComputer[43] and NamSor API. These two tools were selected to try to assure the best inference possible, since a person is only considered from a gender if both tools appoint that. Also the two tools were utilized in Qiu et al. [32] with the same process described here. GenderComputer, to infer the gender related to a name, receives two parameters: the name; and the person's country of origin (because the same name can be common to different genres in different countries). If the value related to the country of origin is not sent, the tool will try to infer it and, therefore, may be less accurate [45]. Namsor is a data mining tool that tries to predict a person's origin details based on his or her name only [1].

For this work, since the data collected did not have that information, the classification was performed taking into account only the name. In order to reach a consensus between the tools (Namsor and GenderComputer), we considered only the records in which both tools classified the contributor as the same gender - male or female - and, in case of lack of consensus, the "Unkown" value was given. Figure 2 details the process of gender inference in this work.

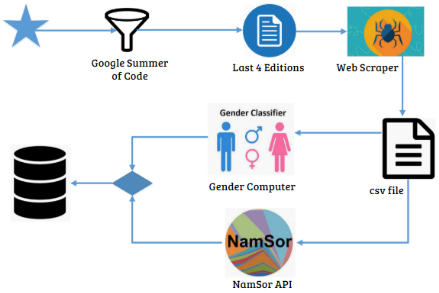

Figure 2: Gender Inference Process

The statistical analysis was performed utilizing the result of Gender Inference process, which is also a CSV file. The content of this file was analyzed utilizing the programming language R, a language and environment for statistical computing and graphics that provides a wide variety of statistical and graphical techniques, and is highly extensible [33]. After this analysis, Python was once

[1] https://www.namsor.com/

more utilized to implement the Apriori algorithm to find the association rules within the data set. The details regarding the steps of the Data Mining process are described in the further sections. All the coding efforts used in this process, as well as the data needed to execute the code and replicate the results, are available in the project repository in dropbox[2].

3.2 CRISP-DM

The project described here uses the CRISP-DM reference model. This model defines a set of sequential steps to guide data mining and enables the mining process to be fast, reliable and with greater management control [11]. CRISP-DM involves six phases: business understanding, data understanding, data preparation, modeling, evaluation, and implementation, as shown in the Figure 3. The adoption of these phases assists in defining the flows to be used to execute the mining project [24].

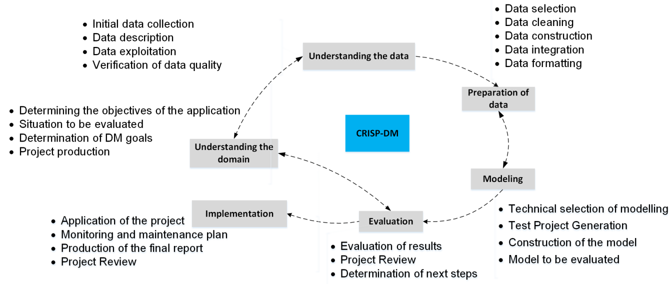
• Initial data collection
• Data description
• Data exploitation
• Verification of data quality
• Data selection
• Data cleaning
• Data construction
• Data integration
• Data formatting
Understanding the data
Preparation of data
• Determining the objectives of the application
• Situation to be evaluated
• Determination of DM goals
• Project production
CRISP-DM
Understanding the domain
Modeling
Implementation
Evaluation
• Application of the project
• Monitoring and maintenance plan
• Production of the final report
• Project Review
• Technical selection of modelling
• Test Project Generation
• Construction of the model
• Model to be evaluated
• Evaluation of results
• Project Review
• Determination of next steps

Figure 3: Phases of the CRISP-DM Methodology [9]

Business Understanding: Corresponds to the understanding of project goals and requirements from the business perspective, followed by the conversion of this knowledge into the definition of the mining problem and the preliminary project plan to achieve the objectives. In this phase, to represent the open source community the GSoC database is being used, which is made up of data from the last four editions and contains several attributes related to the submitted projects, such as: company, developer, gender, project url, project title, developer type, year, main technology and main topic of the project. It is important to highlight that the program (GSoC) does not have its own project repository and, therefore, a good part of these projects are in tools such as Github. The last four editions were selected because the official website only keeps the complete archive data since 2016. The purpose of this research is to discover if the available attributes will be related in algorithms for the definition of association rules.

Data Understanding: it consists of the initial data collection, familiarization with the data and identification of possible problems with the data quality, aiming to discover the first insights about the data or to detect interesting information in the subsets for hypothesis formation. In this phase, statistical analysis was performed on the data in order to facilitate understanding. The results of this analysis are described in the Results Section.

Data Preparation: performs data preparation by covering all activities to construct the final data set obtained from the initial raw data. The data set variables were prepared to generate the models used in the next phase. The main idea of data preparation in this context is to build formatted data sets for use. The Google Summer of Code (GSoC) contributors data was collected through an web scraper (a tool built to retrieve data from websites) that took data from the last four editions of GSoC (2016, 2017, 2018 and 2019). Since the original data doesn't identify the gender, in order to carry out the Gender Inference of all the contributors, was used a gender discovering process called GenderComputer developed by Vasilescu et al. [43] with the NamSor APINamSor API)[3], as already commented. As the data collected didn't had information regarding the country of origin of the contributors (information that gender discovering tools usually ask for), the classification was performed taking into account only the name.

After getting all the related data, some data were discarded due to contextual irrelevancy or with the goal to anonymize the data set. The discarded data were: the name of the developers; the code repository URL of the project; the company name; and the project name. The result of this process is a data set with the same amount of rows, but a reduced amount of columns. Finally, the resultant table were sliced in two different data sets: the first with the men's data and the second one with women's data. This processing was planned due to the goal of the project.

Modeling: define models and modeling techniques that will be applied to accurately organize data. In this article, we choose the already described Apriori Association Rule Algorithm. The results related to it are presented in the Analysis of Results section.

Evaluation: evaluates the model obtained in more detail and reviews the steps for the construction of the final model in order to ensure that it meets its objectives properly. Here we are utilizing the lift value of the association rules to evaluate them;

Deployment: consolidates the knowledge discovered with the created model. The purpose of the model is to increase knowledge about the data and present it in a way that can be useful. This phase can be as simple as generating a final report, or as complex as implementing a repetitive data mining process in a given organization. For this project, all CRISP-DM phases will be used, except the deployment phase, which will be proposed in the future works section.

4 ANALYSIS OF RESULTS

4.1 Statistical Analysis

4.1.1 Gender Related Statistics. The dataset retrieved from GSoC has a total of 12323 rows. From those 2419 participated in 2016; 3307 in 2017; 3081 in 2018; and 3516 in 2019. From the total amount of the four years, 26.065% of the participants were from the female gender. The percentage by year is that were 42.17% (2016), 22.56% (2017), 21.78% (2018) and 22.04% (2019). Interestingly, in 2016, the number of women participants in GSoC was quite expressive – more than 42% – and in the last three years there has been a drop of more than 20% in the number of women participants and the participation of women in 2017, 2018 and 2019 maintained an average of 22.12%. This result is quite expressive, since the average participation of women in open source projects is 1-5% [26]. In addition, women are very

[2] https://bit.ly/2EpzhLk

[3] https://www.namsor.com/

underrepresented on GitHub projects in relation to programming activities [44]. Robles et al. [35] conducted a survey with GitHub contributors and found that only 10% of contributors are female. Izquierdo et al. [23] also stated that women are underrepresented in the OpenStack community. Thus, we can conclude from the last 4 editions of the GSoC that, although women are underrepresented, the percentage of women participants in GSoC is higher than in other communities. One of the reasons for this representativeness may be due to the various existing mentoring programs with the aim of attracting women to participate in the training programs.

Year	Male	Female	Unknown	Total
2016	1329	1020	70	2419
2017	2407	746	154	3307
2018	2244	671	166	3081
2019	2543	775	198	3516
Total	8523	3212	588	12323

Table 1: Participants By Gender And Year

The complete relationship between number of participants by gender (male, female and unknown) and year can be found at the Table 1. From the data displayed on this table is also possible to realize that, even tough we have a considerable amount of people with unknown gender (588), they represent only 4.77% of the total amount.

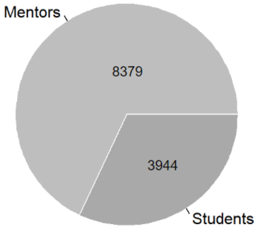

Figure 4: Mentors and Students

4.1.2 Other Statistics. Among the 12323 rows in the data set, 8379 are from mentors and only 3944 are from students. Making explicit that the biggest part of the GSoC participants are mentoring others. The Figure 4 shows a pie chart that makes easier the comparison of this numbers. When it comes to mentoring percentage the genders similarly are almost equivalent since 69.33% of the women are mentors and 67.63% of men are mentors.

With the purpose of interpreting the technologies within the GSoC data, all the values were ordered in a list with the absolute frequency of utilization. This qualitative variable has a total of 131 occurrences. After that, the data was sorted, and the top 5 occurrences were selected as input for the bar plot of the Figure 5. The graph shows that the most frequent technologies in the projects of the referred database are, in this order, "Python", "Java", "C++", "C" and "JavaScript". The chart also communicates that the database has more than 2500 projects with "Python", while the second most used technology (Java) has less than 1500.

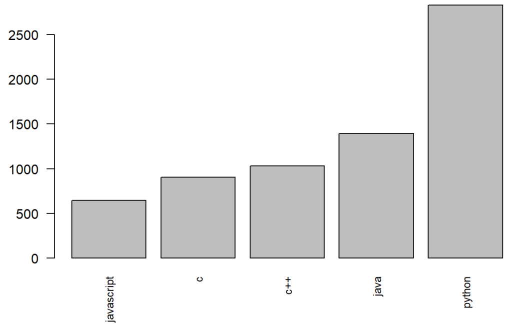

Figure 5: Most Utilized Technologies

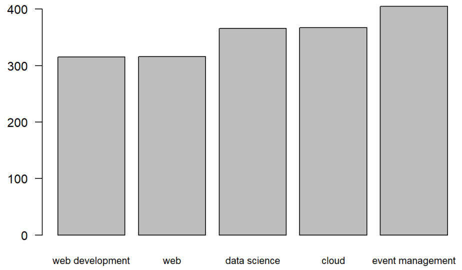

Figure 6: Most Common Topics

In the process of understanding the main topics of the project within the GSoC data, all values were ordered in a list with the absolute frequency of use. This qualitative variable has a total of 345 occurrences. After that the data was sorted and the top 5 occurrences were selected as input for the bar plot of the Figure 6.

The graph shows that the "Event Management" is the most common topic, followed by "Cloud"; "Data Science"; "Web"; and "Web Development". The graph also shows that the distribution of topics by project has higher levels of equality than the distribution of technologies shown in the Figure 5.

4.2 Data Mining

As told before, this project aims to find association between gender characteristics and other project characteristics through the mining of Association Rules. The CRISP-DM reference model was followed within the methodology of this work, hence, part of the data mining process is already described in the Methodology section. The current section shows and explains details about the discovered Association Rules. The two aforementioned data sets (with men and women) were processed with Apriori Algorithm under the same parameters, being them:

- **Support**: being Support(A, B) equals to the amount of tuples containing both A and B divided by the total of tuples, the support value was defined as "0,005" meaning that the rule has to appear at least 62 times in the data set with 12323 items. The value "0,05" was also tested, but returned no rules; "0,04" returned only one rule; "0,03" returned 3 rules; and "0,01" returned 15 rules;

- **Confidence**: being Confidence(A, B) equals to the amount of tuples containing both A and B divided by the amount of tuples containing A, the confidence value was defined as "0,7" meaning that a rule is only considered reliable if the frequency of occurrences of B where A occurs is higher or equal "0,7";
- **Length**: being the length value the minimum amount of items to be associated in a rule, the length value was defined as "2" meaning that rules with only one item are not considered; and
- **Lift**: being lift(A, B) equals to Confidence(A, B) divided by Support(A, B), the lift value was defined as "3,0" meaning that the association rule was picked as a strong rule only if the chance of B occurring between items with A is, at least, three times bigger than in the other items.

The processing of the algorithm resulted in 37 association rules for women's data set and 39 for the men data set. Between those 22 rules are unique for women's data set; 24 are unique for men's data set; and 15 rules are not unique, hence belong to both data sets. The top unique rules generated for women's data set are shown in the Table 2.

Those were classified according to the "lift" of the rule. The main rules in this context are the associations between ios technologies and mobile applications (both with students and mentors); the technology "qt" (pronounced "cute", it's a C++ framework where you can build software without writing code through actions like drag and drop) and desktop applications (mainly with women that are mentors) and the association between R project and data science (both with students and mentors).

Rule	Confidence	Lift
ios >mobile applications	1	94.47
mobile applications >ios	1	94.47
ios; mentor >mobile applications	1	94.47
mentor; mobile applications >ios	1	94.47
mentor; qt >desktop applications	0.7	77.53
desktop applications >qt	1	74.69
mentor; desktop applications >qt	1	74.69
student; data science >r-project	0.76	41.37
mentor; r-project >data science	1	41.17
student; r-project >data science	1	41.17

Table 2: Top Association Rules - Female Data Set

From this result set 2 is notable that there are some two way bindings (when there's two rules where the rule body and the rule head exchange the places). It's also notable that in the top 10 rules, there's no programming languages and only one programming framework: qt. R project is a software environment for statistical computing and graphics, so it's not considered here as a synonym for R (the programming language of the environment).

The top unique rules generated in the men data set are shown in the Table 3. As well as the rules for the female data set, those were classified according to the "lift" of the rule. It's possible to realize that the mean lift is lower in this data set (being 72.85 for the first data set and 17.35 for this one).

Rule	Confidence	Lift
physics >c/c++	1	54.63
physics; mentor >c/c++	1	54.63
electronic voting >Scala	0.99	43.04
logic >Scala	0.97	43.03
electronic voting; mentor >Scala	1	43.03
logic; mentor >Scala	1	43.03
kubernetes >cloud	0.95	37.88
artificial intelligence; mentor >JavaScript	0.71	14.96
mentor; machine translation >c++	1	11.8
privacy >c++	0.82	9.74

Table 3: Top Association Rules - Male Data Set

The main rules in this context are the associations between the topic of "physics" and the programming languages "c/c++"; between works with "logic" as the main topic and the programming language Scala; the association between "kubernetes" and "cloud"; between artificial intelligence and JavaScript; and between the programming language C++ and the topics of "machine translation" and "privacy". It's also possible to realize from these association rules that the ones generated from the men data set have more terms directly related to programming languages and logic.

Rule	Confidence	Lift
kernel >c	0.81	37.84
data science >r-project	1	37.84
r-project >data science	0.74	28.13
r >data science	1	14.37
kernel >c	1	14.36
kernel; mentor >c	1	11.94
machine translation >C++	1	11.93
office suite >c++	1	8.73
creative coding >java	1	4.18
event management >python	1	4.18

Table 4: Top Association Rules - Common

The top association rules that are present as strong rules in both data sets (men and women) are presented in Table 4. It's important to realize that this rules are not a result of an analysis of the whole data set ignoring the gender (this approach would not be accurate since it could show rules that are valid only for one gender); these are rules that could be found in both data sets and the values for "Confidence" and "Lift" presented are the mean between these values in both result sets.

5 LIMITATIONS AND THREATS TO VALIDITY

One threat to the validity of our results is that 588 items in the data set couldn't get a gender following our gender inference process.

Thus, this data (without gender) was discarded in the association rules mining. Adding to that, is also a threat to the validity the accuracy of the already described process of gender inference; the only way to be sure about this information would be questioning each participant about his gender, but it was considered impracticable for this work. In addition, the gender inference process is more accurate if the country of origin is informed [45] and as we do not have this information in our context, this is also be considered a threat to validity.

The data set is limited and full of rare classes. Given that, a big part of the association rules are very obvious and trivial (as the association between "ios" and "mobile applications", for example). The last presented threat to validity is the limited size of the database to be utilized in the association rules mining. To have more accuracy, it would be necessary to have a bigger database.

6 CONCLUSION

This work presented a comprehensive research about the participation of men and women in the area of Information and Communications Technology (ICT) through data extracted from the last four editions of GSoC. The goal of this project is to find Association Rules between gender characteristics and coding using the Apriori Algorithm and to analyze statistical data related to the context. In the course of the building of this work, the main definitions about Gender Diversity, Google Summer of Code, Data Mining and Association Rules were spelled out to give the necessary background and, also in background section, some Related Works were shown. The method consisted of using the CRISP-DM reference model and planning data mining according to six phases: business understanding; data understanding; data preparation; modeling; evaluation; and deployment.

From the results of this work - considering also all the threats to validity previously presented - it's possible conclude that the women representativeness in GSoC (a global program focused on bringing more developers into open source software development) is decreasing in recent years. Besides that, a total of 61 association rules were mined, being 22 of them found only in the data set with women, 24 found only with the men, and 15 applicable to both sets. An important conclusion that can be drawn from our work is that analyzing and comparing the rule sets mined for men and women, it's possible to realize that men are more related to terms (topics and technologies) that are directly linked to coding/programming.

When it comes to the most utilized technologies, we have "Python", "Java", "C++", "C" and "JavaScript" in the top. Analyzing the association rules mined from the female data set it's possible to realize that none of these technologies are inside a rule (as head or body) and the same does not happen within the association rules extracted from the male data set - it contains rules with "C++" and "JavaScript". Besides that, the common association rule set we have rules with "C"; "C++"; "Java"; and "Python". This shows that the main utilized technologies for men and women are similar, but, in general, male persons utilize more programming languages.

When it comes to the most common project topics, we have "Event Management", "Web", "Web Development", "Data Science" and "Cloud" in the top. Analyzing the association rules mined from the female data set it's possible to see that only "Data Science" appear between the top ten association rules. For the data set with male data, only "Cloud" from the most common project topic is included between the top ten association rules. This can represent how diverse the project topics of the database are, but not necessarily has something related to gender. The high lift values in the unique rules can represent that these rules are trivial. It's also possible to see that some very near rules are unique in both sets; it can represent that the association is, in fact, valid in both contexts.

As a future work it is proposed to identify more topics by work - not only one - and create association rules between them with the goal to identify trending related topics in IT open-source community. Furthermore, it's also valid to propose a phase to clean the data regarding Technologies and Topics before the analysis. With that relations like "python 3" and "python" would be classified as the same technology - something that's not happening in the current work.

In addition, the insertion of the cultural element in the analysis is also a proposed future work. This analysis will be possible after mining the ethnicity and/or the nationality of people with techniques like the one developed by Treeratpituk and Giles [40]. This can help identifying which cultures have better women representativeness. Furthermore, as a last proposed future work, based on the presented results two questions are raised to be researched: 1. why has female representativeness not increased in recent years; and 2. why men are more directly linked to coding.

REFERENCES

[1] Monica Babes-Vroman and Thu D. Nguyen. 2020. Gender Diversity in Computer Science at a Large Research University. *CoRR* abs/2004.13760 (2020), 1–25. arXiv:2004.13760 https://arxiv.org/abs/2004.13760

[2] Joseph Beck, Min Chi, and Ryan S. Baker. 2017. Workshop proposal: deep learning for educational data mining. In *Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017, June 25-28, 2017*. International Educational Data Mining Society (IEDMS), Wuhan, Hubei, China. http://educationaldatamining.org/EDM2017/proc_files/papers/paper_15.pdf

[3] Kelly Blincoe, Olga Springer, and Michal R. Wróbel. 2019. Perceptions of Gender Diversity's Impact on Mood in Software Development Teams. *IEEE Software* 36, 5 (2019), 51–56. https://doi.org/10.1109/MS.2019.2917428

[4] Amiangshu Bosu and Kazi Zakia Sultana. 2019. Diversity and Inclusion in Open Source Software (OSS) Projects: Where Do We Stand?. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2019*. IEEE, Recife, Brazil, 1–11. https://doi.org/10.1109/ESEM.2019.8870179

[5] Carmen Botella, Silvia Rueda, Emilia López-Iñesta, and Paula Marzal. 2019. Gender Diversity in STEM Disciplines: A Multiple Factor Problem. *Entropy* 21, 1 (2019), 30.

[6] Carmen Botella, Silvia Rueda, Emilia López-Iñesta, and Paula Marzal. 2019. Gender Diversity in STEM Disciplines: A Multiple Factor Problem. *Entropy* 21, 1 (jan 2019), 30. https://doi.org/10.3390/e21010030

[7] Danielle Brown and Melonie Parker. 2019. Google diversity annual report 2019. https://diversity.google/ (Date last accessed 16-April-2019).

[8] Cássio Oliveira Camilo and João Carlos da Silva. 2009. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)* 1, 1 (2009), 1–29.

[9] Edna Dias Canedo, Rhandy Rafhael de Carvalho, Heloise Acco Tives Leão, Pedro Henrique Teixeira Costa, and Márcio Vinicius Okimoto. 2019. How the Academics Qualification Influence the Students Learning Development. In *43rd IEEE Annual Computer Software and Applications Conference*, Vol. 1. IEEE, Milwaukee, USA, 336–345. https://doi.org/10.1109/COMPSAC.2019.00056

[10] Edna Dias Canedo, Giovanni Almeida Santos, Fabiana Freitas Mendes, Elaine Venson, and Rejane Maria da Costa Figueiredo. 2018. Why there is still few women in Engineering? A perspective from female students and professors in an Engineering campus. In *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, San Jose, USA, 1–8. https://doi.org/10.1109/fie.2018.8659171

[11] P Chapman, J Clinton, R Kerber, T Khabaza, T Reinartz, C Shearer, and R Wirth. 2000. CRISP-DM 1.0 Step-by-step data mining guide. Crisp DM Consortium (Updated 2010)(1999).

[12] Sapna Cheryan, Allison Master, and Andrew N. Meltzoff. 2015. Cultural stereotypes as gatekeepers: increasing girlsâ€™ interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology* 6 (February 2015), 1–49. https://doi.org/10.3389/fpsyg.2015.00049

[13] LinkedIn Corporate Communications. 2019. 2019 LinkedIn Workforce Diversity Report. https://news.linkedin.com/2019/January/our-2019-diversity-report (Date last accessed 27-May-2020).

[14] Crummy. 2020. Beautiful Soap. https://www.crummy.com/software/BeautifulSoup/ (Date last accessed 31-May-2020).

[15] Leandro Augusto da Silva, Sarajane Marques Peres, and Clodis Boscarioli. 2017. *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil, Rio de Janeiro, Brazil. 296 pages.

[16] Rosanna Durruthy. 2018. 2018 LinkedIn Workforce Diversity Report. https://careers.linkedin.com/diversity-and-inclusion/workforce-diversity-report (Date last accessed 16-April-2019).

[17] Stack Exchange. 2018. Computer Science Educators Stack Exchange. https://cseducators.stackexchange.com/ (Date last accessed 31-May-2020).

[18] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17, 3 (1996), 37–54. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

[19] Feng Feng, Junghoo Cho, Witold Pedrycz, Hamido Fujita, and Tutut Herawan. 2016. Soft set based association rule mining. *Knowl. Based Syst.* 111 (2016), 268–282. https://doi.org/10.1016/j.knosys.2016.08.020

[20] Dalibor Fiala. 2005. *Web Mining and Its Applications to Researchers Support.* Ph.D. Dissertation. University of West Bohemia, Pilsen, Czech Republic.

[21] Google. 2019. Google Summer of Code. https://summerofcode.withgoogle.com/ (Date last accessed 16-April-2019).

[22] Sander Hoogendoorn, Hessel Oosterbeek, and Mirjam van Praag. 2013. The Impact of Gender Diversity on the Performance of Business Teams: Evidence from a Field Experiment. *Management Science* 59, 7 (2013), 1514–1528. https://doi.org/10.1287/mnsc.1120.1674

[23] Daniel Izquierdo, Nicole Huesman, Alexander Serebrenik, and Gregorio Robles. 2019. OpenStack Gender Diversity Report. *IEEE Software* 36, 1 (2019), 28–33.

[24] Pankush Kalgotra and Ramesh Sharda. 2016. Progression analysis of signals: Extending CRISP-DM to stream analytics. In *2016 IEEE International Conference on Big Data*. IEEE Computer Society, Washington, USA, 2880–2885. https://doi.org/10.1109/BigData.2016.7840937

[25] Stylianos Kampakis. 2020. *The Decision Maker's Handbook to Data Science: A Guide for Non-Technical Executives, Managers, and Founders.* Apress, London, England. 166 pages. https://doi.org/10.1007/978-1-4842-5494-3

[26] Amanda Lee and Jeffrey C. Carver. 2019. FLOSS participants' perceptions about gender and inclusiveness: a survey. In *ICSE*. IEEE / ACM, 10.1109/ICSE.2019.00077, 677–687.

[27] Ghadeer Mobasher, Ahmed Shawish, and Osman Ibrahim. 2017. Educational Data Mining Rule based Recommender Systems. In *CSEDU 2017 - Proceedings of the 9th International Conference on Computer Supported Education*, Vol. 1. SciTePress, Porto, Portugal, 292–299. https://doi.org/10.5220/0006290902920299

[28] Leon Moonen, Stefano Di Alesio, David Binkley, and Thomas Rolfsnes. 2016. Practical Guidelines for Change Recommendation Using Association Rule Mining. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering* (Singapore, Singapore) *(ASE 2016)*. Association for Computing Machinery, New York, NY, USA, 732–743. https://doi.org/10.1145/2970276.2970327

[29] Sukanya Kannan Moudgalya, Kathryn M. Rich, Aman Yadav, and Matthew J. Koehler. 2019. Computer Science Educators Stack Exchange: Perceptions of Equity and Gender Diversity in Computer Science. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education, SIGCSE*. ACM, Minneapolis, USA, 1197–1203. https://doi.org/10.1145/3287324.3287365

[30] Noelia Olmedo-Torre, Fermin Sanchez Carracedo, M. Nuria Salan Ballesteros, David Lopez, Antoni Perez-Poch, and Mireia Lopez-Beltran. 2018. Do Female Motives for Enrolling Vary According to STEM Profile? *IEEE Transactions on Education* 61, 4 (Nov. 2018), 289–297. https://doi.org/10.1109/te.2018.2820643

[31] Pandas. 2020. About pandas. https://pandas.pydata.org/about/index.html (Date last accessed 31-May-2020).

[32] Huilian Sophie Qiu, Alexander Nolte, Anita Brown, Alexander Serebrenik, and Bogdan Vasilescu. 2019. Going farther together: the impact of social capital on sustained participation in open source. In *Proceedings of the 41st International Conference on Software Engineering, ICSE*. IEEE / ACM, Montreal, Canada, 688–699. https://doi.org/10.1109/ICSE.2019.00078

[33] R-Project. 2020. What is R? https://www.r-project.org/about.html (Date last accessed 31-May-2020).

[34] Penny Rheingans, Erica D'Eramo, Crystal Diaz-Espinoza, and Danyelle Ireland. 2018. A Model for Increasing Gender Diversity in Technology. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education, SIGCSE*. ACM, Baltimore, USA, 459–464. https://doi.org/10.1145/3159450.3159533

[35] Gregorio Robles, Laura Arjona Reina, Alexander Serebrenik, Bogdan Vasilescu, and Jesús M. González-Barahona. 2014. FLOSS 2013: a survey dataset about free software contributors: challenges for curating, sharing, and combining. In *MSR*. ACM, 10.1145/2597073.2597129, 396–399.

[36] Rafael Santos. 2009. Conceitos de Mineração de dados na web. *XV Simpósio Brasileiro de Sistemas Multimídia e Web, VI Simpósio Brasileiro de Sistemas Colaborativos–Anais, MM Teixeira, CAC Teixeira, FAM Trinta, e P. PM Farias, Eds* 1, 1 (2009), 81–124.

[37] Jefferson De Oliveira Silva, Igor Wiese, Daniel M. Germán, Christoph Treude, Marco Aurélio Gerosa, and Igor Steinmacher. 2020. Google summer of code: Student motivations and contributions. *J. Syst. Softw.* 162 (2020), 1–40. https://doi.org/10.1016/j.jss.2019.110487

[38] Jefferson O. Silva, Igor S. Wiese, Igor Steinmacher, and Marco A. Gerosa. 2017. Students' Engagement in Open Source Projects: An Analysis of Google Summer of Code. In *Proceedings of the 31st Brazilian Symposium on Software Engineering* (Fortaleza, CE, Brazil) *(SBES'17)*. Association for Computing Machinery, New York, NY, USA, 224–233. https://doi.org/10.1145/3131151.3131156

[39] Erik H. Trainer, Chalalai Chaihirunkarn, Arun Kalyanasundaram, and James D. Herbsleb. 2014. Community Code Engagements: Summer of Code & Hackathons for Community Building in Scientific Software. In *GROUP*. ACM, 10.1145/2660398.2660420, 111–121.

[40] Pucktada Treeratpituk and C. Lee Giles. 2012. Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, Toronto, Canada, 1141–1147. http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5180

[41] Marie E. Vachovsky, Grace Wu, Sorathan Chaturapruek, Olga Russakovsky, Richard Sommer, and Fei-Fei Li. 2016. Toward More Gender Diversity in CS through an Artificial Intelligence Summer Program for High School Girls. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education, Memphis, TN, USA, March 02 - 05, 2016* (Memphis, Tennessee, USA) *(SIGCSE '16)*. ACM, https://dl.acm.org/doi/10.1145/2839509.2844620, 303–308. https://doi.org/10.1145/2839509.2844620

[42] Raimundo Claudio Vasconcelos, Antonio Justiniano Moraes Neto, and Lúcio Teles. 2018. PROPOSTA DE UM MODELO DE MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAR A COLABORAÇÃO ENTRE ESTUDANTES DA EAD. *CIET: EnPED* 2018, 1 (2018), 1–17.

[43] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2014. Gender, Representation and Online Participation: A Quantitative Study. *Interacting with Computers* 26, 5 (2014), 488–511.

[44] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G. J. van den Brand, Alexander Serebrenik, Premkumar T. Devanbu, and Vladimir Filkov. 2015. Gender and Tenure Diversity in GitHub Teams. In *CHI*. ACM, 10.1145/2702123.2702549, 3789–3798.

[45] Bogdan Vasilescu, Alexander Serebrenik, and Vladimir Filkov. 2015. A Data Set for Social Diversity Studies of GitHub Teams. In *12th IEEE/ACM Working Conference on Mining Software Repositories, MSR 2015, Florence, Italy, May 16-17, 2015*. IEEE Computer Society, https://ieeexplore.ieee.org/document/7180131, 514–517. https://doi.org/10.1109/MSR.2015.77

[46] Jennifer Wang, Hai Hong, Jason Ravitz, and Marielena Ivory. 2015. Gender Differences in Factors Influencing Pursuit of Computer Science and Related Fields. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education* (Vilnius, Lithuania) *(ITiCSE '15)*. Association for Computing Machinery, New York, NY, USA, 117–122. https://doi.org/10.1145/2729094.2742611