A machine learning examination of women's leadership effectiveness in software development processes

Sâmara Ahyeska Alves Ferreira[1], Danielli Araújo Lima[1]

[1]Laboratório de Inteligência Computacional, Robótica e Otimização (LICRO)
Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro
(IFTM) Campus Patrocínio, MG, Brasil

samara.ferreira@estudante.iftm.edu.br, danielli@iftm.edu.br

Abstract. *The inclusion of women in information technology companies and software development processes is vital for fostering diverse perspectives and innovative problem-solving. We analyzed* 793 *instances representing globally distributed software development teams, aiming to show that female-led teams outperform male-led ones. Through descriptive statistics and Welch's t-test, we confirmed this hypothesis. Using a decision tree with only three inputs—female leadership presence, total team members, and female team members—we achieved* 76.79% *accuracy, significantly reducing computational time compared to using all* 85 *dataset attributes. This approach also informs recommendation systems for assembling development teams, emphasizing the value of gender diversity in enhancing team dynamics and solutions in the tech industry.*

1. Introduction

Software engineering, a field within computer science, involves activities ranging from designing and creating software to evaluating and deploying it [Naseer et al. 2020]. Additionally, Information Technology (IT) has become indispensable across various industries, playing a crucial role in streamlining processes and cutting service delivery times [Alves et al. 2021]. Many organizations and universities choose team-based strategies that involve multiple participants, often with teams distributed globally. Furthermore, the digitization of businesses has become a global trend, spurred by factors like government regulations, market competition, and the pursuit of operational efficiency.

Facing challenges such as refining project concepts, meeting deadlines, and resolving conflicts among team members can strain relationships and necessitate team restructuring [Petkovic et al. 2016]. Particularly in academia, students often face obstacles in forming teams, communicating effectively, meeting deadlines, and adhering to project guidelines [Soares et al. 2021]. However, as technology evolves, professionals must continually adapt and enhance their skills to navigate through novel and sometimes daunting situations [Rodrigues et al. 2022].

In the realm of Software Engineering, successful software implementation and development hinge upon effective team management, whether in the academic context or in companies [Petkovic et al. 2014]. However, it's notable that a disproportionately small number of software development teams are led by women, reflecting the broader gender disparity prevalent in the field of information technology [Lima et al. 2021, Ferreira et al. 2019]. Nonetheless, as the IT landscape evolves, there's a growing recognition of the importance of gender diversity, particularly within sectors like software and application development processes [Alves et al. 2021, Rodrigues et al. 2022]. Consequently, companies are increasingly championing the inclusion of women in technology roles, acknowledging the value they bring to innovation and team dynamics [Rodrigues et al. 2022, Lima et al. 2021].

An essential area for consideration is the efficacy of teams managed by women. It is imperative to evaluate whether such teams can achieve high performance levels in their designated tasks and pinpoint the key factors contributing to their success [Beghoura 2021]. Thus, this research aims to delve deep into the impact of women in team leadership roles on the success of software development teams operating across various team configurations, including the percentage of women in teams. In our exploration of previous literature, no studies were discovered that delved into the analysis of [Petkovic et al. 2012] dataset or examined the influence of female team leaders on team performance.

Transitioning into our initial analysis, we began by employing descriptive statistics, which was followed by the application of Welch's t-test to validate our hypotheses. Subsequently, this study employs an artificial intelligence (AI), know as machine learning techniques, specifically decision trees, to identify the most relevant attributes pertaining to team compositions led by women and the quantity of women within teams. The decision tree is pivotal in establishing a set of rules for recommending an effective team composition, employing a concise array of attributes and minimal rules for decision-making. It is crucial to note that our goal is not to outperform other studies in accuracy, given their use of the entire dataset. Rather, we aim for strong accuracy by focusing on female attributes in team assembly, offering insights for future research.

2. Methodology

This study utilizes real-world data to examine teams led by women in software development phases and their success factors. It employs comprehensive data analysis with Artificial Intelligence (AI) techniques, aligning with applied IT research. This study aims to analyze various software development teams, assessing the influence of female leadership on team performance and identifying optimal team compositions. It is descriptive research with explanatory aims, delving into the nature of relationships between variables.

Our study conducts experiments with variables using a dataset in `.xlsx` format within the KNIME Analytics Platform [Dornelas and Lima 2023]. The research involves 40 steps, incorporating machine learning techniques, particularly the C4.5 decision tree model (DT) [Shafer et al. 1996]. It follows an experimental framework, defining the object of study, selecting influential variables, and establishing control rules, particularly focusing on the impact of female leadership in software development.

As materials, a dataset sourced from the UCI Machine Learning Repository[1] is utilized, comprising team activity measurements and machine learning results from student teams in Software Engineering projects across various universities. This quantitative approach characterizes the study as a single case study, with the dataset holding promise for predicting student learning from teamwork activities.

The dataset includes Team Activity Measures (TAMs) and outcomes from 74 student teams in software engineering classes across multiple semesters at three universities

[1] SETAP Dataset: Data for Software Engineering Teamwork Assessment in Education Setting `https://archive.ics.uci.edu/dataset/393/data+for+software+engineering+teamwork+assessment+in+education+setting`.

[Petkovic et al. 2016]. It focuses on students' engagement and performance in final class projects. Each team, consisting of 5-6 students, is graded as `A` or `F` based on process adherence and product quality. The data, gathered through weekly timecards, instructor observations, and tool usage logs, is aggregated into TAMs [Petkovic et al. 2016]. These TAMs are valuable for classification tasks in Computer Science, particularly for analyzing sequential, time-series data. The dataset's mix of integer and real features makes it suitable for various analytical approaches to understand and improve team dynamics and project outcomes in software engineering education [Petkovic et al. 2016].

3. Results

In this section, we present three types of analysis. The first experiment involves a statistical descriptive analysis to examine the statistics of manually selected attributes. The second experiment focuses on our hypothesis test to verify whether teams led by women exhibit significantly better performance in mean compared to teams led by men. Finally, we utilize decision tree learning with three different attribute compositions. Our objective is to demonstrate that using attributes related to women can result in good accuracy that can be used to create process developement teams in future.

3.1. Descriptive statistics

Table 1 provides statistical summaries for various parameters related to team performance and composition. The `teamMemberCount` column indicates that the teams consist of a minimum of 3 members and a maximum of 7, with an average team size of approximately 5.19 members. The low standard deviation suggests that team sizes generally cluster closely around the mean. Regarding the `femaleTeamMembersPercent`, the data reveals that the percentage of female team members ranges from 0% to approximately 83.33%. The mean value of approximately 0.176 suggests that, on average, female members constitute around 17.6% of the team composition. In terms of `teamLeadGender`, the mean value of approximately 0.193 indicates that, on average, about 19.3% of the teams are led by females. The positive skewness suggests that there are relatively fewer female-led teams compared to male-led ones. Similarly, for `teamDistribution`, the mean value of approximately 0.201 indicates that, on average, around 20.1% of the teams have a specific distribution. The positive skewness indicates that a higher proportion of teams fall into this category compared to those with other distributions.

Table 1. Descriptive statistics of manually selected attributes used in the decision tree learning algorithm.

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum
teamMemberCount	3	7	5.189155	1.197444	1.433872	-0.54983	-0.52666	4115
femaleTeamMembersPercent	0	0.8333	0.176106	0.167126	0.027931	1.247057	2.342405	139.6522
teamLeadGender	0	1	0.192938	0.394854	0.15591	1.559251	0.432347	153
teamDistribution	0	1	0.200504	0.40063	0.160505	1.498902	0.247323	159
teamMemberResponseCount	1	84	28.24464	17.12803	293.3694	0.883988	0.255952	22398
commitCount	0	783	115.0883	121.1723	14682.73	1.954473	5.684651	91265
issueCount	0	12	1.715006	1.886151	3.557565	1.852734	4.598062	1360
onTimeIssueCount	0	10	1.356873	1.532479	2.348491	1.665572	3.827164	1076
lateIssueCount	0	4	0.358134	0.689285	0.475114	2.200428	5.216313	284
timeInterval	1	11	6.090794	3.146158	9.898312	-0.03243	-1.21469	4830

The `teamMemberResponseCount` column reveals that the average number of responses per team member is approximately 28.24, with a wide range of responses ranging from 1 to 84. The high variance suggests significant variability in response rates across teams. The `commitCount` column shows that teams make an average of approximately 115 commits, with a wide range of commit counts from 0 to 783. The high kurtosis indicates a heavy-tailed distribution, suggesting that some teams make an unusually large number of commits. Regarding `issueCount` and `onTimeIssueCount`, the mean values are approximately 1.72 and 1.36, respectively, indicating the average number of issues and on-time issues per team. The high kurtosis for `issueCount` suggests a heavy-tailed distribution, while the positive skewness for `onTimeIssueCount` indicates that the majority of teams have fewer on-time issues. Lastly, the `timeInterval` column suggests that the average time interval between events is approximately 6.09 units, with a standard deviation of approximately 3.15. The negative skewness indicates that the distribution of time intervals is skewed to the left, with a longer tail on the left-hand side.

3.2. Hypothesis test

The summary statistics provide a comprehensive overview of the performance metrics for teams led by women (Group G_1) compared to those led by men (Group G_2) using the AAT Bioquest[2]. Group G_1 has a higher average performance score ($\mu_1 = 0.8562$) compared to G_2 ($\mu_2 = 0.6234$). The 95% Confidence Intervals for G_1 and G_2 are $(0.912, 0.661)$, indicating a high probability that the true mean scores lie within these ranges. Both groups have similar median scores of 1. G_1 shows less variability (variance 0.1239, standard deviation 0.352) than G_2. Performance scores for both groups range from 0 to 1. With a larger sample size in G_1 (640) compared to G_2 (153), these statistics provide insights into the performance of teams led by women and men in software development.

The Welch's t-test was conducted to compare the performance of teams led by female and male leaders using G_1 and G_2, respectively. We formulated the following hypotheses:

Null Hypothesis (H_0): There is no significant difference between the mean scores of the G_1 teams led by women and G_2 teams led by men ($\mu_1 = \mu_2$).

Alternative Hypothesis (H_a): There is a significant difference between the mean scores of the G_1 teams led by women and G_2 teams led by men ($\mu_1 \neq \mu_2$).

The resulting p-value of $6.0494e - 11$ indicates a highly significant difference in performance between the two groups. With a calculated t-statistic of 6.7838 and degrees of freedom (df) of 306.1791, the test underscores the robustness of the observed disparity. The 95% Confidence Interval for the difference in means $(0.1653, 0.3003)$ further supports the conclusion that teams led by female leaders exhibit significantly higher performance compared to those led by male leaders. Therefore, we reject the null hypothesis (H_0), indicating that there is indeed a significant difference between the performance of teams led by female and male leaders.

3.3. Machine learning

Table 2 presents the results of the decision tree algorithm applied to different sets of attributes in the dataset. Each row represents a specific set of attributes used in the model,

[2] Quest Graph T-Test Calculator. AAT Bioquest, Inc., 25 Mar. 2024, `https://www.aatbio.com/tools/one-two-sample-independent-paired-student-t-test-calculator`.

while the columns provide associated performance analysis. Attribute sets comprise three categories: “All 85 dataset Attributes”, “10 Manually selected Attributes” (displayed in Table 1), and “3 selected Female Attributes” (the first three elements shown in Table 1).

Table 2. Output of decision tree algorithm applied into dataset.

Columns considered	Class	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen’s kappa
85 features	F	190	63	467	73	0.722433	0.750988	0.722433	0.881132	0.736434	0.828499	0.609402
	A	467	73	190	63	0.881132	0.864815	0.881132	0.722433	0.872897		
10 features	F	194	66	464	69	0.737643	0.746154	0.737643	0.875472	0.741874	0.829760	0.614881
	A	464	69	194	66	0.875472	0.870544	0.875472	0.737643	0.873001		
3 features	F	153	74	456	110	0.581749	0.674009	0.581749	0.860377	0.624490	0.767970	0.457915
	A	456	110	153	74	0.860377	0.805654	0.860377	0.581749	0.832117		

Performance metrics include true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), sensitivity, precision, sensitivity, specificity, F-measure, accuracy, and Cohen’s Kappa.

These analyses evaluate the model’s performance in accurately assessing instances within each class. Results show varying impacts of different attribute sets on model performance, with notable differences observed. For instance, the “All Attributes” set achieved an accuracy of approximately 0.8284, while the “Female Attributes” set yielded 0.7679, and the highest accuracy came from manual attribute filtering at 0.8297. Even with just three female attributes, the significance remains, highlighting gender’s importance in classifying team performance. This aligns with hypothesis test findings, emphasizing gender diversity’s significance.

Using a small number of attributes, it is possible to determine an effective team configuration based on female attributes, considering team size. While our goal is not to surpass previous works, our paper is concerned with isolating the most significant variables to reduce computational processing. When using 10 features, we achieved an overall accuracy of 82.976%, and when using only female attributes, the accuracy was 76.797%, surpassing [Petkovic et al. 2016], who achieved only 70% accuracy. Gender acts as a significant guide, enabling the model to discern patterns and make informed classifications.

4. Conclusions and future work

This study examined performance gaps between teams led by women and men in software development, using hypothesis testing and machine learning. Findings showed significant differences, with female-led teams outperforming and exhibiting less variability. The Welch t-test confirmed these results, rejecting the null hypothesis. Machine learning analysis deepened the understanding of team dynamics, particularly highlighting the influence of gender diversity. Notably, attribute sets with female attributes achieved high accuracy rates, emphasizing gender’s importance in team performance. Future research could explore gender diversity’s influence on team performance, including communication dynamics and leadership styles, to gain deeper insights.

Acknowledgments

Author SAAF is particularly grateful for the scholarship support from the CNPq under the PIBITI. Additionally, author DAL thanks the FAPEMIG, CAPES and IFTM for their financial support.

References

Alves, L. M., Nascimento, S. M., and Silva, V. M. (2021). Investigando a participação das mulheres nas áreas de teste e qualidade de software. In *Anais do XV Women in Information Technology*, pages 305–309. SBC.

Beghoura, M. A. (2021). Software engineering teamwork data understanding using an embedded feature selection. *International Journal of Performability Engineering*, 17(5):464.

Dornelas, R. S. and Lima, D. A. (2023). Correlation filters in machine learning algorithms to select demographic and individual features for autism spectrum disorder diagnosis. *Journal of Data Science and Intelligent Systems*, 1(2):105–127.

Ferreira, M. E., Lima, D. A., and Silva, A. (2019). Data analysis for robotics and programming project evaluation involving female students participation. In *2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE)*, pages 417–422. IEEE.

Lima, D. A., Ferreira, M. E. A., and Silva, A. F. F. (2021). Machine learning and data visualization to evaluate a robotics and programming project targeted for women. *Journal of Intelligent & Robotic Systems*, 103(1):4.

Naseer, M., Zhang, W., and Zhu, W. (2020). Early prediction of a team performance in the initial assessment phases of a software project for sustainable software engineering education. *Sustainability*, 12(11):4663.

Petkovic, D., Okada, K., Sosnick, M., Iyer, A., Zhu, S., Todtenhoefer, R., and Huang, S. (2012). Work in progress: a machine learning approach for assessment and prediction of teamwork effectiveness in software engineering education. In *2012 frontiers in education conference proceedings*, pages 1–3. IEEE.

Petkovic, D., Sosnick-Pérez, M., Huang, S., Todtenhoefer, R., Okada, K., Arora, S., Sreenivasen, R., Flores, L., and Dubey, S. (2014). Setap: Software engineering teamwork assessment and prediction using machine learning. In *2014 IEEE frontiers in education conference (FIE) proceedings*, pages 1–8. IEEE.

Petkovic, D., Sosnick-Pérez, M., Okada, K., Todtenhoefer, R., Huang, S., Miglani, N., and Vigil, A. (2016). Using the random forest classifier to assess and predict student learning of software engineering teamwork. In *2016 IEEE frontiers in education conference (FIE)*, pages 1–7. IEEE.

Rodrigues, M. E. M., Maia, A. M. A., Rocha, M. d. S., de Oliveira, L. M. C., and Marques, A. B. (2022). Desenvolvimento de soft skills durante a atuação no projeto meninas digitais do vale: achados de uma retrospectiva. In *Anais do XVI Women in Information Technology*, pages 34–44. SBC.

Shafer, J., Agrawal, R., Mehta, M., et al. (1996). Sprint: A scalable parallel classifier for data mining. In *Vldb*, volume 96, pages 544–555. Citeseer.

Soares, A. L., Ferreira, M. E. A., and Lima, D. A. (2021). Experience report and data visualization to evaluate a game programming project aimed for girls using scratch tool. In *Anais do XXVII Workshop de Informática na Escola*, pages 43–52. SBC.