electronics

MDPI

Article

A Data-Driven Approach to Team Formation in Software Engineering Based on Personality Traits

Jan Vasiljević and Dejan Lavbič *

Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia; jv1721@student.uni-lj.si
* Correspondence: dejan.lavbic@fri.uni-lj.si

Abstract: Collaboration among individuals with diverse skills and personalities is crucial to producing high-quality software. The success of any software project depends on the team's cohesive functionality and mutual complementation. This study introduces a data-centric methodology for forming Software Engineering (SE) teams centred around personality traits. Our study analysed data from an SE course where 157 students in 31 teams worked through four project phases and were evaluated based on deliverables and instructor feedback. Using the Five-Factor Model (FFM) and a variety of statistical tests, we determined that teams with higher levels of extraversion and conscientiousness, and lower neuroticism, consistently performed better. We examined team members' interactions and developed a predictive model using extreme gradient boosting. The model achieved a 74% accuracy rate in predicting inter-member satisfaction rankings. Through graphical explainability, the model underscored incompatibilities among members, notably those with differing levels of extraversion. Based on our findings, we introduce a team formation algorithm using Simulated Annealing (SA) built upon the insights derived from our predictive model and additional heuristics.

Keywords: team formation; personality traits; Software Engineering; data-driven approach; Simulated Annealing

Citation: Vasiljević, J.; Lavbič, D. A Data-Driven Approach to Team Formation in Software Engineering Based on Personality Traits. *Electronics* **2024**, *13*, 178. https://doi.org/10.3390/electronics13010178

Academic Editors: Mohamed Wiem Mkaouer and Chunping Li

Received: 30 November 2023
Revised: 22 December 2023
Accepted: 28 December 2023
Published: 30 December 2023

1. Introduction

Team formation (TF) is critical in many domains, including business, sports, and academia [1,2]. In Software Engineering (SE), however, TF assumes a unique and pivotal role. SE, an engineering discipline encompassing all aspects of software production [3], is deeply rooted in collaboration. The complexity of software projects often necessitates a team of engineers with diverse specializations, making these teams' efficiency, communication, and synergy crucial for project success. This study aims to bridge the gap in TF by integrating psychological metrics with machine learning techniques to optimize team composition in SE.

Historically, TF has relied heavily on empirical metrics, but recent trends have shifted towards incorporating psychological aspects to enhance team dynamics and performance [4]. The dynamics of how team members interact and engage with each other are crucial in SE. Research has consistently shown a strong correlation between positive team dynamics and the success of high-performing teams [5–8]. However, integrating psychological traits into automated TF systems, especially in SE, requires further exploration.

This gap is particularly evident when considering the subtler yet influential factor of team members' psychological and personality traits. These traits can profoundly impact how individuals approach problems, interact with colleagues, and respond to stress or success. There has been a growing interest in using personality frameworks, such as the Big Five Personality Traits, to build technically proficient and psychologically compatible teams.

The following section reviews the existing literature to identify gaps and limitations and inform our proposed solution. Subsequent sections outline our methodologies, describe

the data collection process, detail the models used, and provide an initial overview of the proposed algorithm. This is followed by a presentation of our results, encompassing an overview of the collected data, the trained model, and the newly developed TF algorithm. The paper concludes with a discussion of our findings, summarizing our contributions, acknowledging limitations, and suggesting guidelines for future research.

2. Related Work

The interplay between individual personality traits and team performance, particularly in collaborative and skill-intensive environments, has been widely studied. A hierarchical model developed by [9] categorizes key personality facets vital for team performance, offering a detailed understanding of how broad personality traits correlate with specific team requirements. This model utilizes specific facets of higher-level dimensions, such as adjustment, flexibility, and dependability, to predict team adaptability, interpersonal cohesion, and decision making.

A study by [10] investigated the impact of personality traits on the performance of product design teams comprising undergraduate engineering students. The findings indicated a significant positive correlation among conscientiousness, openness, and team performance. In the study, groups of three students were tasked with building a bridge using limited resources, emphasizing the profound influence of personality traits over other factors, including cognitive ability and demographic diversity.

Research by [11] delved into the predictive capacity of conscientiousness facets within engineering student project teams over a 6.5-month-long task. The primary conclusion was that conscientiousness effectively forecasted team performance. However, it was noted that other traits, such as agreeableness, extraversion, and neuroticism, did not have a significant predictive impact on team performance.

In the realm of SE, ref. [12] assessed the impact of personality traits on SE team effectiveness using the Myers–Briggs Type Indicator (MBTI). They highlighted the significant role of personality clashes in software project failures and pointed out a gender-based variance in MBTI traits among programmers. Suggestions for optimal trait balances for male and female team members were presented.

The research study [13] undertook a systematic literature review from 1970 to 2010 centring on individual personalities in SE. The review offers an extensive overview of the field's current understanding, notably highlighting the diversity of results.

The study [14] provided direct evidence of the impact of specific personality traits in an SE context, emphasizing the significance of extraversion in promoting effective team dynamics. Additionally, the study revealed that openness to experience positively correlates with team performance, not team climate. However, the limited sample size of respondents might limit the generalizability of these findings.

In their research, ref. [15] mapped the job requirements of various SE roles to the Big Five Personality Traits, suggesting specific personality traits beneficial for different SE roles. This mapping includes the need for extraversion and agreeableness in system analysts; openness and conscientiousness in software testers; and a combination of extraversion, openness, and agreeableness in programmers.

The authors in [16] extended the application of personality traits in team dynamics to an educational perspective, focusing on team dynamics within the context of student programming projects. This study addresses the gap between academic projects and real-world software development, offering insights into managing and facilitating student projects that closely mimic professional environments.

In reviewing the foundational work of [9], we see a detailed model of personality facets relevant to team performance. Although theoretically robust, the model does not offer a direct pathway for practical application in SE team assembly.

Similarly, the study by [12] proposed guidelines for balancing teams based on members' traits and gender but did not provide a practical framework for applying these findings.

The predominant theme in these studies is an in-depth theoretical exploration of personality impacts and extensive data analysis. However, there is a noticeable lack of tools or methods designed to practically apply these insights to forming SE teams. Furthermore, there is minimal focus on the automated prediction of team members' interactions and performance. While guidelines are provided, their practical implementation in real-world scenarios is constrained due to the extensive manual labour required for their application.

3. Proposed Solution

In our work, we address the gap in TF by introducing a data-driven approach tailored for SE. Recognizing the limitations in practical applications highlighted by previous studies, our approach builds upon their theoretical foundations. We focus on developing practical solutions that are informed by methodological advancements. Our key contributions include the following:

- Data collection in a controlled SE environment: Our research involved collecting data on key aspects, such as personality traits, team performance, and inter-member satisfaction. The collection was carried out in a controlled SE environment, which helped ensure that our data were accurate and relevant to typical SE settings.
- Predictive model for inter-member satisfaction: We developed a predictive model with our collected data. This model is aimed at understanding and forecasting inter-member satisfaction in SE teams. It considers various factors, such as personality traits and work contributions, providing valuable insights into team dynamics and cohesion.
- Team formation algorithm: Our algorithm integrates the predictive model directly, utilizing its forecasts and various heuristics to aid in forming effective SE teams. Designed with practical implementation in mind, it provides a systematic method for real-world team assembly. The algorithm functions by accepting participants with associated personality traits as input and outputting strategically formed teams.

To conduct our research, we relied on established methods. Figure 1 visually summarizes our approach, illustrating these methods' integration in addition to data collection and a high-level overview of the proposed TF algorithm. Section 4.1 further describes the process of data acquisition. We used the Five-Factor Model (FFM) (see details in Section 4.2) to assess personality traits, the Extreme Gradient Boosting algorithm (XGBoost) (see details in Section 4.3) for developing our predictive model, and Simulated Annealing (SA) (see details in Section 4.4) for our TF algorithm (see details in Section 4.5). Python was our primary tool for data analysis.

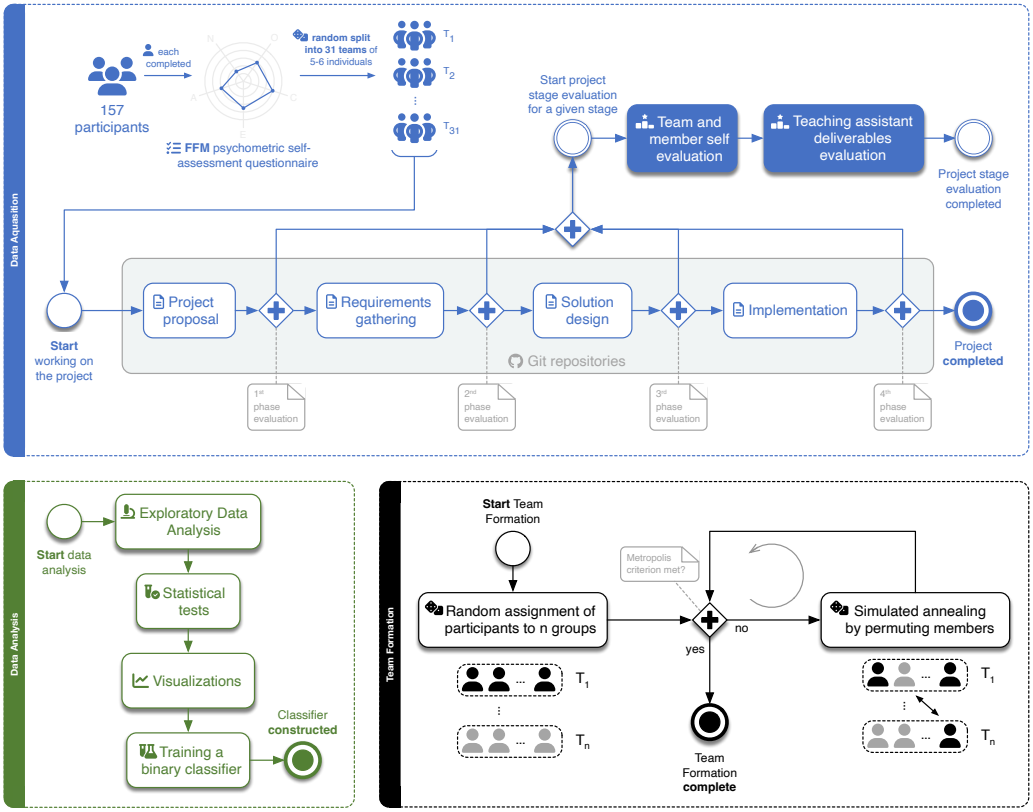


Figure 1. Overview of the proposed approach.

4. Materials and Methods

4.1. Data Acquisition and Description

Data for this study were sourced from the mandatory course Software Engineering at the Faculty of Computer and Information Science, University of Ljubljana. The study encompassed 157 third-year undergraduates, divided into 31 teams of 5 to 6, whose contributions were anonymized for analysis. The course's curriculum was segmented into four stages: project proposal, requirement gathering, solution design, and implementation. The project phases spanned 3 months, each lasting 2–3 weeks.

Before the project began, students completed a 41-item psychometric questionnaire based on the Five-Factor Model, utilizing a 5-point Likert scale to gauge personality traits for insights into team dynamics. Following each phase, surveys were collected to assess team satisfaction, focusing on performance, communication, and individual contributions.

Version control was managed using Git within a GitHub organization, allowing for the extraction of commit data, including the number of commits and lines of code. After each stage, teaching assistants evaluated the teams' deliverables using a 0–100 scale.

4.2. Five-Factor Model

The Five-Factor Model, also known as the OCEAN model, is a prominent psychological framework for evaluating human personality along five key dimensions: openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N) [17]. In our study, participants underwent a standardized test to measure these dimensions. Each questionnaire item aligns with a specific dimension and carries an "influence value", denoted by $I(q)$, reflecting the question's framing.

Particularly, when a question negatively correlates with the trait it assesses, the corresponding response, $R(q)$, is inverted to represent the trait accurately. To quantify the dimension score (S) for a set of related questions (Q), we calculate S as the weighted sum of the responses ($R(q)$) and their respective influence values ($I(q)$), normalized to the

maximum possible score for that set, max_score(Q). Mathematically, this is represented as $S = \frac{\sum_{q \in Q} R(q) \times I(q)}{\text{max\_score}(Q)}$.

Despite some criticism, such as not fully accounting for all variances in human personality [18] or the lack of complete independence among variables [19], the FFM is a reliable and valid model for measuring personality traits in SE domains [20].

4.3. Xgboost and Shapley Additive EXplanations

To create a predictive model focusing on the assessment of inter-member satisfaction, we employed the Extreme Gradient Boosting algorithm, widely recognized as XGBoost. XGBoost is an open-source machine learning framework known for its computational efficiency and robust performance metrics. As a type of ensemble learning, it is an optimized version of gradient-boosted decision trees tailored for speed and accuracy. Noteworthy features of XGBoost include its built-in capacity to handle missing data, support for parallel processing, and its versatility in addressing a diverse range of predictive problems [21].

In addition to model prediction, interpretability is crucial when dealing with large "black box" machine learning models. This research uses Shapley Additive EXplanations (SHAP) to provide a detailed understanding of feature influence on predictions [22].

4.4. Simulated Annealing

To address the NP-hard challenge of TF, we employed Simulated Annealing [23], a heuristic optimization algorithm. The algorithm was designed to optimize the composition of student teams by using heuristics from the FFM and the XGBoost predictive model. SA initiates with a solution (S_{init}) and, in each iteration, generates a new candidate solution (S') from the neighbourhood ($N(S)$) of S. The acceptance of S' over S is dictated by the *Metropolis* criterion, which considers the change in objective function $\Delta E = E(S') - E(S)$. A solution S' is accepted if $\Delta E \leq 0$ or if a randomly generated number between 0 and 1 is less than $\exp(-\Delta E/T)$, where T is the current temperature. The algorithm iterates for a fixed number of iterations (N_{stop}) and returns the most optimized solution found.

4.5. Overview of the Team Formation Algorithm

The TF algorithm is designed to input a collection of participants and output optimally assembled teams. It primarily processes participant data characterized using the FFM personality assessment. The algorithm necessitates specifying the quantity and size of the teams, ensuring that the total aligns with the student count. Additionally, it involves setting weights for the objective function, which aims to balance the sub-scores across teams while maximizing them overall. The algorithm operates through three primary processes:

1. Simulated Annealing: This stage involves generating new candidate teams and progressively lowering the system's temperature.
2. Satisfaction evaluation: Utilizing the XGBoost predictive model, this step assesses the satisfaction levels within each group. This metric is incorporated into the objective function.
3. Scoring and decision making: Each team's score is computed using the designated objective function. Based on the criteria set by the Simulated Annealing process, the algorithm decides whether to accept or reject the new team formations.

The output is a structured array of teams, each with assigned members. The algorithm maintains equitable scores across teams while maximizing individual member satisfaction. For an in-depth discussion on the objective function and additional details, refer to Section 5.5. The process is graphically represented in Figure 2.

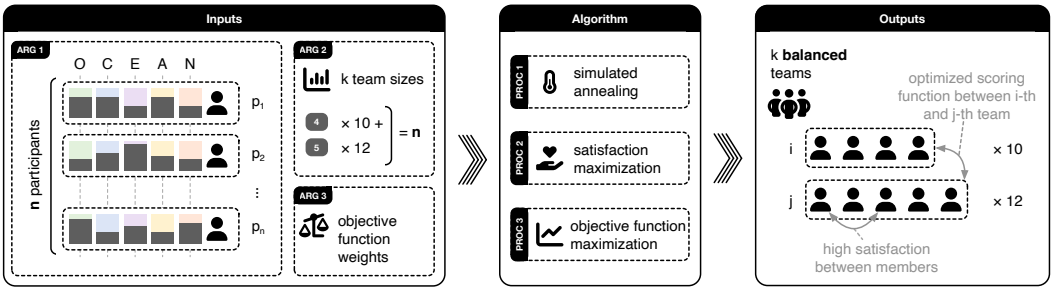


Figure 2. Team formation algorithm inputs, algorithm processes and outputs.

5. Results

5.1. Data Overview

5.1.1. Phase Results

After each project phase, students were evaluated and graded by TAs. The grading criteria were based on quantifiable aspects, such as the volume of work completed, project deliverables, and the overall quality of the work. Rather than team grades, individual grades were assigned on a scale from 0 to 100. These grades were communicated to each student during a meeting with the TA and their respective team. The distribution of these grades across different phases is illustrated in the histograms shown in Figure 3.

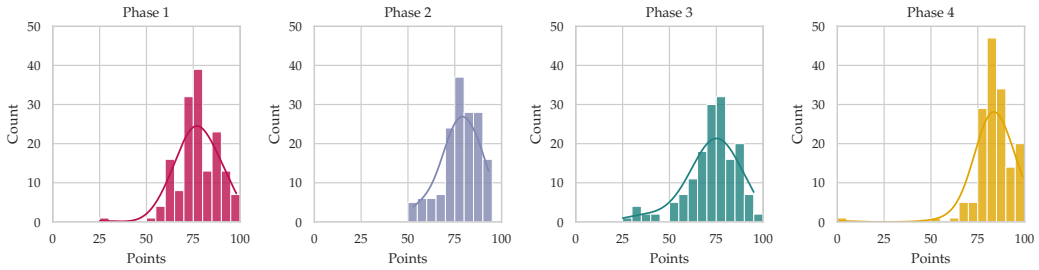


Figure 3. Histograms of grades assigned to individual students in each phase.

The average grades for the first two phases were relatively consistent, with $\mu_1 = 77.7$ for the first phase and $\mu_2 = 77.8$ for the second. However, the third phase proved more challenging, reflected by a drop in the average grade to $\mu_3 = 73.0$. The average grade increased again in the fourth phase to $\mu_4 = 84.0$. This pattern aligns with student feedback, which indicated that the third and fourth phases were the most difficult. The improved performance in the fourth phase can be attributed to its focus on project implementation, a task with which the students were the most familiar.

5.1.2. FFM Data

Aggregate data of the distribution of students are depicted in Table 1. We compared the results to a study on the effect of teammate personality on team production [24]. Our observed average for conscientiousness stands at 0.69, exceeding the 0.59 reported in the referenced study. This disparity can be attributable to our participants being in their third academic year, suggesting a more developed work ethic than first-year undergraduates. Our dataset displayed slightly higher averages for openness (0.67) and agreeableness (0.61). In contrast, extraversion averaged at 0.54, a significant deviation from the 0.75 cited in the comparative study. Neuroticism showcased consistency across both datasets, with our average resting at 0.37 compared to 0.38 in the reference study.

Table 1. Means and standard deviations of the FFM personality traits.

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Mean	0.67	0.69	0.55	0.61	0.38
Std	0.15	0.14	0.14	0.13	0.16

To measure the linear relationship of each pair of traits, we constructed a correlation matrix utilizing the PCCs. The values in this matrix span the interval $[-1, 1]$. A value approaching 1 indicates a strong positive relationship, while a value nearing -1 signifies a pronounced negative relationship. Conversely, values proximate to 0 indicate minimal to non-existent correlation.

Key insights derived from the correlation matrix, presented in Table 2, are as follows:

- Conscientiousness and neuroticism: A moderate negative correlation ($r = -0.33$) was observed between conscientiousness and neuroticism. This suggests that individuals scoring higher in conscientiousness tend to exhibit fewer neurotic traits.
- Extraversion and openness: The data indicated a positive correlation ($r = 0.20$) between extraversion and openness.
- Conscientiousness and extraversion: A notable correlation existed between conscientiousness and extraversion, evidenced by a coefficient of $r = 0.23$.

The observed relationships in the first two points are consistent with prior research [25,26]. The third point, however, presents an ambiguous relationship, which could be attributed to the limited sample size or the specific demographic characteristics of the student population studied.

Table 2. Correlation matrix of FFM dimensions among students. Symbols denote significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	O	C	E	A	N
O (openness)	-				
C (conscientiousness)	0.18 *	-			
E (extraversion)	0.20 *	0.23 **	-		
A (agreeableness)	0.18 *	0.00	−0.01	-	
N (neuroticism)	−0.07	−0.33 ***	−0.17 *	−0.08	-

5.1.3. Repository Commit Data

Git data were collected from each team's repository, into which students committed their code and documentation. The initial dataset comprised 59,305 commits harvested from the git repositories of the student teams. Several filtering steps were implemented to ensure relevance and accuracy. Commits outside the project deadlines were excluded to align with the grading period. Additionally, commits with erroneous timestamps—attributable to misconfigured git clients—were removed. Non-source-code elements like tool-generated directories (for example `node_modules` generated by a package manager) and build artefacts, which do not represent a developer's effort, were also omitted.

Furthermore, commits that merely consisted of minified CSS files within documentation directories—often a byproduct of wireframe creation—were considered non-essential and were excluded. A noticeable pattern of commits with large but equal counts of line additions and deletions was attributed to code formatting tools. To address this, commits were filtered using a heuristic that targeted those with line modifications exceeding 50, ensuring that trivial style changes did not inflate the data.

After these preprocessing steps, the commit count was refined to 28,466, representing approximately 48% of the initial volume. A visual representation of the filtered commit activity across teams is shown in Figure 4. The vertical dashed orange lines represent the deadlines for individual phases.

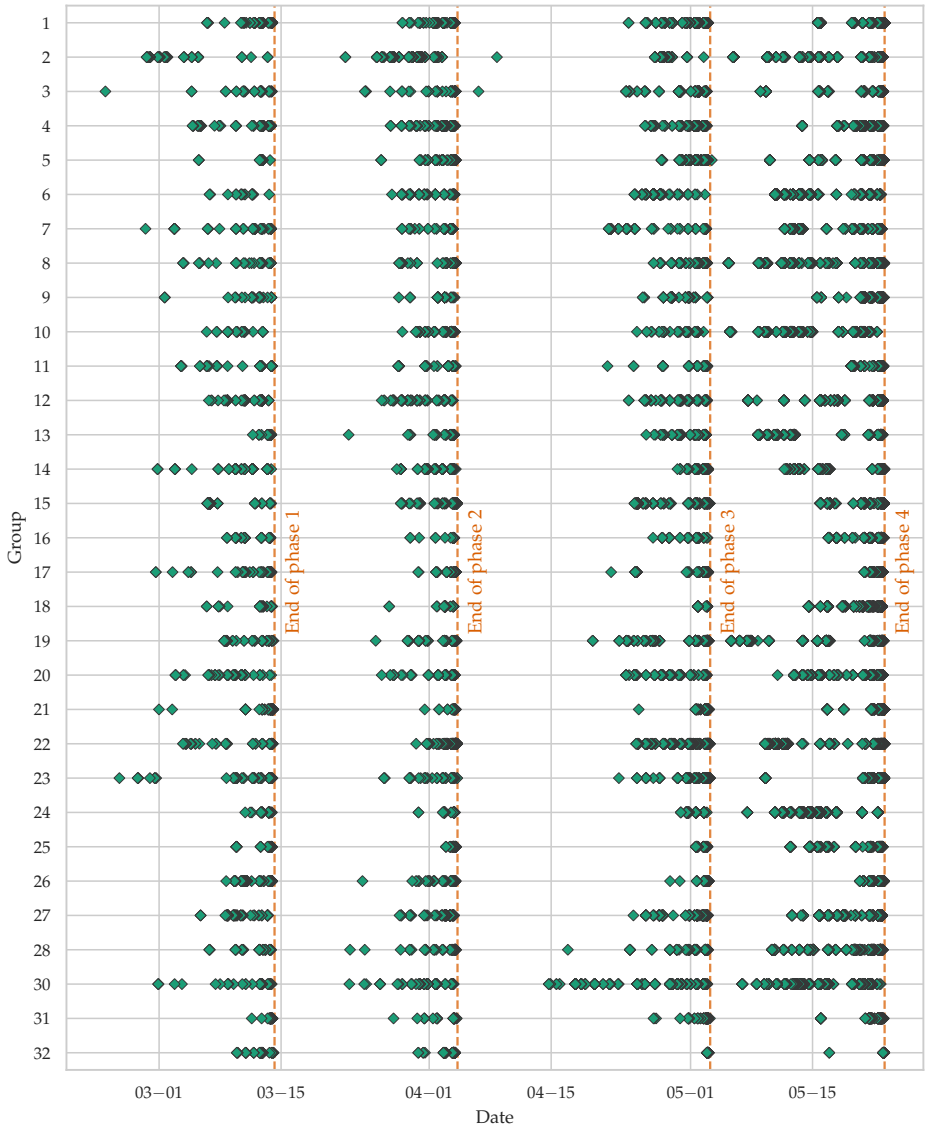


Figure 4. Stripplot of filtered-out commits of various groups. The vertical dashed orange lines represent the deadlines for individual phases.

5.2. Team Performance and FFM Dimensions

We define team performance as the arithmetic mean of grades in each project phase. Using the median performance as the separation criterion, teams were classified as high-performing or low-performing teams, as depicted in Figure 5. The variations in FFM dimensions between these groups were statistically significant and aligned with previous research, confirming the following:

1. Openness and agreeableness do not have a meaningful effect on team performance. Contrary to the assumption that teams with open or agreeable members perform better, the data do not support this claim.
2. Conscientiousness has been confirmed in past research [24] to have a positive association with team performance. Teams with organized, dependable, and hardworking members tend to outperform others.
3. Neuroticism hurts how teams fare, which aligns with previous findings [10]. High neuroticism in individuals can lead to challenges within the team, affecting overall effectiveness.
4. Extraversion plays a complex role. Although it has advantages and disadvantages, it positively influences team performance in our dataset. Since computer science stu-

dents may generally be more introverted, extroverted members can be advantageous. Their engagement can counterbalance any issue, like dominating conversations.

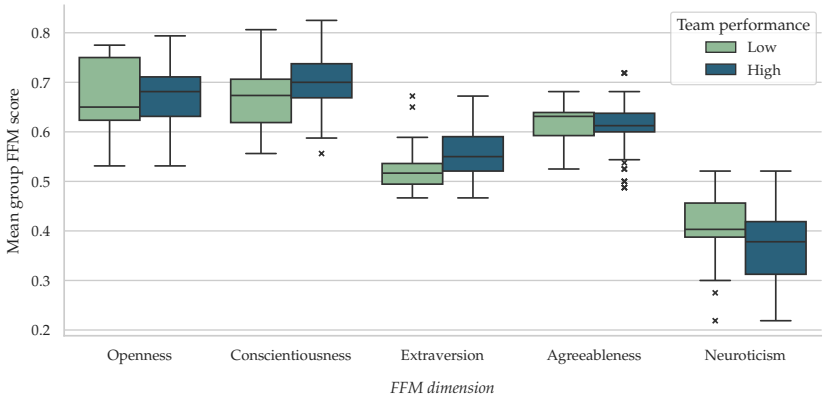


Figure 5. Box plots with outliers of the FFM dimensions.

Before assessing the significance of our findings, we checked if the conditions for a parametric t-test were met. Levene's test confirmed that the variances between the two groups were equal. However, the Shapiro–Wilk test revealed that only the extraversion and conscientiousness dimensions were normally distributed. Considering these points, we opted for the non-parametric Mann–Whitney U test to examine the mean differences across the dimensions. The test highlighted significant differences between high-performing and low-performing teams in the following dimensions: conscientiousness ($p = 0.005$), extraversion ($p = 0.0001$), and neuroticism ($p = 0.004$).

5.3. Analysis of Team Satisfaction Metrics

In evaluating team satisfaction during post-project phases, two distinct metrics were employed. The first, a rating scale, $R_1 \in [-2, 2]$, measured individual satisfaction levels, with -2 indicating very low satisfaction and 2 denoting high satisfaction with team members' contributions. Each team member received a non-unique score from their peers. The second metric, a ranking system, $R_2 \in [1, \text{len(teams)}]$, assigned each member a unique rank, with 1 representing the highest satisfaction level.

Team satisfaction was quantified as the normalized sum of all inter-member satisfaction scores, calculated using the formula

$$T = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{s_{ij}}{n-1}}{n} \tag{1}$$

where s_{ij} is the satisfaction score assigned by member i to member j, which could be either R_1 or R_2, and n is the total number of team members. The condition $j \neq i$ excludes self-assessment scores.

Using R_1 as s_{ij}, team satisfaction was calculated for each project phase, and the results are depicted in Figure 6. Phases 2 and 4 showed a notable positive correlation between team satisfaction and performance, with correlation coefficients of $r = 0.39$ and $r = 0.32$. However, Phase 1 exhibited a weak or non-existent relationship, possibly due to initial adjustments among team members. Phase 3's average scores were 5 points lower than the first two phases, leading to reduced satisfaction levels, particularly when outcomes did not meet expectations.

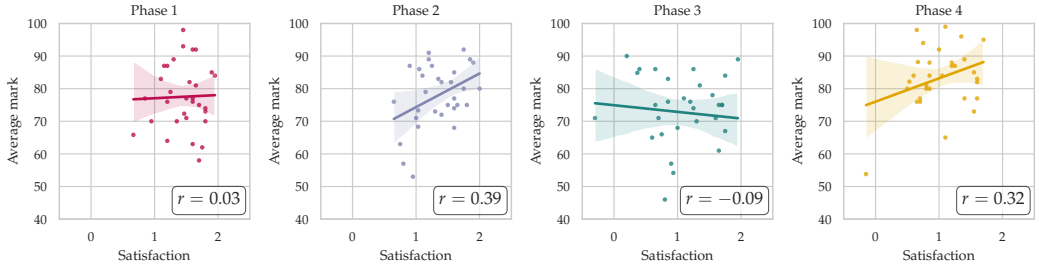


Figure 6. Regression plot of team satisfaction versus team performance for each project phase.

While R_1 correlates with performance, its suitability for our TF method is questionable due to potential data skewness, defined as

$$\text{Skewness (g1)} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{2}$$

Our dataset's skewness was −1.22, indicating strong negative skewness. The Intraclass Correlation Coefficient (ICC) [27] further revealed inconsistencies in the data, possibly due to biased evaluations by students concerned about peer grades.

Consequently, we use the ranked work contribution score (R_2) for further analysis. ICC(3,k) tests on R_2 showed adequate consistency, with 61 out of 108 groups (56.48%) achieving an ICC over 0.5, and 72 groups (66.67%), one exceeding 0.4. These thresholds were based on varied interpretations of the ICC value [28,29]. The chosen threshold is deemed *good* by both standards. This indicates a general agreement on the top contributors, though not unanimous.

A hypothesis for subsequent investigation is that a significant portion of these rankings could be explained by tangible metrics, like commit counts and lines of code, with the remaining variance potentially being linked to distinct personality traits within the teams.

Work Contribution

To effectively measure individual contributions during different project phases, we introduce a metric denoted as $work\_ratio_n$, with n representing the project phase. This metric is derived from Git data, following a specific preprocessing approach.

The calculation of $work\_ratio_n$ involves several steps:

1. The number of lines each member adds (m) is normalized by taking the square root. This adjustment favours smaller, frequent commits instead of larger, sporadic ones. Lines that have been removed are not considered, as they do not reliably indicate work contribution.
2. The normalized line additions for each member (m) are summed up.
3. The total normalized line additions for the team (t) are calculated.
4. The expected work ratio (r) per student is determined as $r = \frac{1}{\text{len(team)}}$. For example, in a five-member team, the expected ratio is 0.2.
5. The individual contribution proportion (p) is computed as $p = \frac{m}{t}$.
6. Finally, the work ratio for phase n is calculated as $work\_ratio_n = \frac{p}{r}$.

The $work\_ratio_n$ metric provides insights into a student's relative contribution:

- A value above 1 indicates contributions exceeding expectations.
- A value below 1 suggests contributions falling short of expectations.
- A value around 1 implies contributions meeting expectations.

The relationship between work contribution and ranking is examined using regression plots (Figure 7). It is important to interpret these findings correctly:

- A lower work contribution suggests that the student contributed more than expected.
- A higher work contribution indicates that the student contributed less than expected.

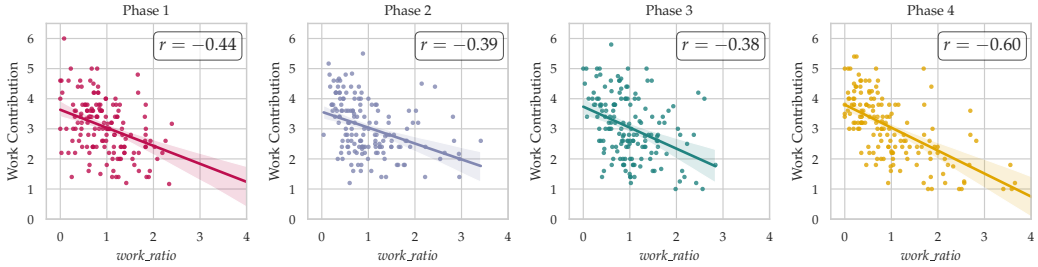


Figure 7. Regression plots of member work contribution ranking versus calculated member work (work_ratio) for each project phase.

The correlation coefficients for each project phase are as follows: $r_1 = -0.44$, $r_2 = -0.39$, $r_3 = -0.38$, and $r_4 = -0.60$. These values indicate a moderate negative correlation between work contribution and ranking. However, the correlation is not strong enough to fully explain the variance in rankings. The following section will explain the remaining variance using gradient boosting.

5.4. Binary Classification of Inter-Member Satisfaction

We want to predict whether two members will successfully collaborate based on their personality traits. We use the R_2 metric as the target variable to achieve this and employ the XGBoost algorithm to create a binary classifier that predicts whether two members will be satisfied with each other. To train this classifier, we use the following features:

- Features 1…5: OCEAN scores of the Ranker member.
- Features 6…10: OCEAN scores of the Target member.
- Feature 11: The work$_n$ ratio of the Target member.

The *Ranker* is designated as the student responsible for assessing the *Target* student. In our approach, rather than both ratios, we solely incorporate the work$_n$ ratio of the *Target*. This decision is based on our strategy to set this variable as a constant during the prediction phase. While adding a second variable might potentially enhance the model's accuracy in the training phase, it is anticipated that it could diminish accuracy in practical application within the TF algorithm.

A dataset of 3108 data points was constructed by pairing each member with another team member. The dataset was split into a training set of 2486 data points and a test set of 622 data points following an 80:20 training–test split. The model was trained using a grid search with cross-validation to identify the best hyperparameters. The optimal model configuration required 300 *estimators*, a *max depth* of 12, and a *learning rate* of 0.1. The model yielded accuracy of 0.74, precision of 0.69, and an ROC-AUC score of 0.79.

To explain how the model works, we utilized SHAP. The beeswarm summary plot, shown in Figure 8, displays the feature importance of the XGBoost model. The most important feature is the work$_n$ ratio of the *Target* member, followed by the *Target* member's OCEAN scores and, finally, the *Ranker* member's OCEAN scores.

The two SHAP waterfall plots in Figure 9 were created to assess how the model works. In the negative classification example, we can see that the low amount of work (5× less than expected) performed by the target student was the main reason for the negative classification but above-average neuroticism also contributed. In the positive classification example, we can see that the high amount of work nudges the classification to be positive. At the same time, the interplay between the *Raters*' and *Target*'s FFM dimensions cancels out.

The dependence plot in Figure 10 suggests that team satisfaction is positively influenced when *Rater* and *Target* have similar extraversion levels. High extraversion in both tends to skew predictions positively, while diverging levels yield adverse outcomes. Similar trends were observed for conscientiousness and neuroticism, reinforcing the model's validity.

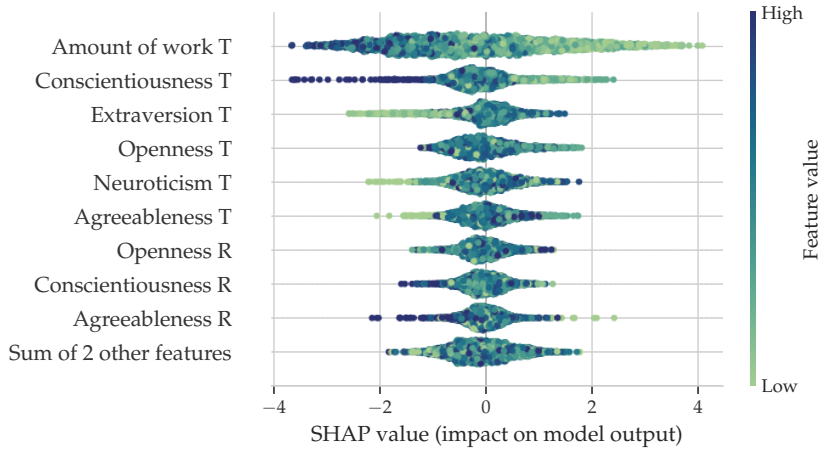
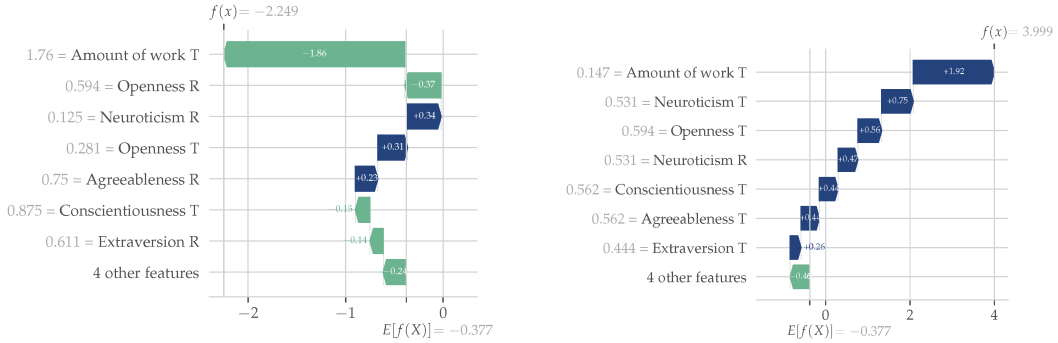


Figure 8. SHAP beeswarm plot of the XGBoost model displaying feature importance for *Rater* (R) and *Target* (T) variables.



(**a**) (**b**)

Figure 9. Waterfall plots of SHAP values for the XGBoost model, illustrating the influence of individual features on the prediction. (**a**) Represents a positive classification. (**b**) Represents a negative classification.

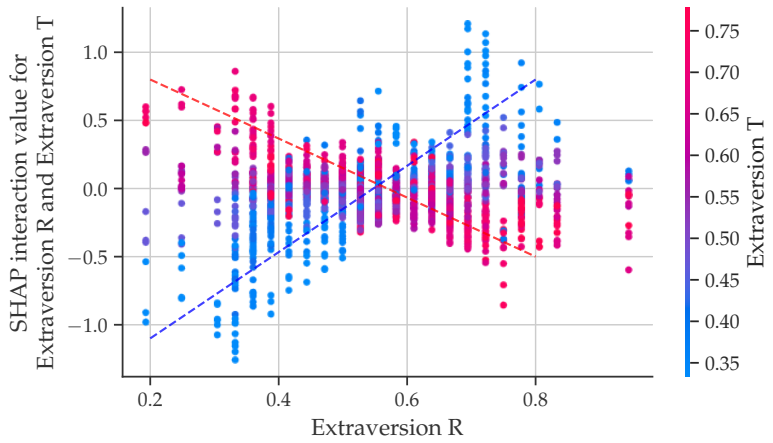


Figure 10. SHAP dependence plot of *Raters'* (R) and *Targets'* (T) extraversion. The dashed lines represent a linear fit applied to the most prominent data points from opposite extremes.

5.5. Team Formation Algorithm Evaluation

Based on prior observations, the following heuristics were established for effective team formation:

- Maximize inter-team R_2 to enhance team performance.

- Ensure teams exhibit high levels of conscientiousness (Q_c) and extraversion (Q_e).
- Aim for low levels of neuroticism (Q_n) within teams.
- Maintain minimal standard deviation across teams for each dimension: Q_c, Q_e, and Q_n.
- Incorporate at least one student with high conscientiousness ($\text{student}_{\max(c)}$) in each team, as suggested by [24]. This approach triggers a beneficial "conscientiousness shock," improving team dynamics.

The Simulated Annealing algorithm was implemented with the following parameters:

1. Initial solution (S_{init}): Teams are formed randomly considering team size distributions, with the constraint that each team includes a student with high conscientiousness. This is achieved by selecting the top N students, where N equals the number of teams. These students are given *locked* status to prevent team swaps.
2. Generation of new solution (S'): A candidate solution is formed by exchanging two students between different teams while ensuring that students with *locked* status remain fixed.
3. Objective function ($E(S)$): The function $E(S) = w_{\text{inter}} \times QT_{\text{inter}} + w_c \times Q_c + w_e \times Q_e + w_n \times \text{NQ}_n$ integrates various dimensions, such as interpersonal satisfaction (QT_{inter}), conscientiousness (Q_c), extraversion (Q_e), and neuroticism (NQ_n). Weights w_{inter}, w_c, w_e, and w_n allow for the adjustment of each dimension's influence.
   Inter-member satisfaction (QT_{inter}): For a team T_k, the average satisfaction (μ_{T_k}) is calculated using the binary classifier's predictions ($p(f_{i,j})$) for each student pair. The global average and standard deviation across all teams are computed to derive $\text{QE}_{\text{inter}} = \mu - \sigma$.
   Scores for Q and NQ: For each dimension, the team average (μ_{T_k}) and global averages (μ and σ) are computed. The resulting scores are defined as $Q = \mu - \sigma$ for Q_c and Q_e, and $NQ = (1 - \mu) - \sigma$ for Q_n.
4. Stopping condition: The algorithm stops after $I_{\text{stop}} = 710N - 1740$ iterations, where N is the number of teams. This formula is based on a linear regression model, with a minimum threshold of 6000 iterations for $N < 10$.

We evaluated the proposed algorithm by comparing it with other methodologies and dimensions, as detailed below:

1. Shuffled: Teams were formed randomly from a predefined set of students.
2. Course: This method represents the teams in their original formation during the course. Although this data point might appear similar to the Shuffled method, it was included separately, as the students' choices could have influenced team compositions.
3. Sorted by conscientiousness: Students were ranked according to their conscientiousness scores and then assigned to teams in a round-robin fashion. This method was employed to compare the SA algorithm's performance with a basic heuristic.

Figure 11 illustrates the operation of the SA algorithm and its comparison with the aforementioned methodologies. The first scenario (left) displays the application of the algorithms on the original dataset, whereas the second scenario (right) applies them to a synthetic dataset. This synthetic dataset was generated by calculating the mean and standard deviation for each dimension of the original dataset and then creating a new dataset with an equivalent number of students and teams. Random values from a normal distribution, based on the mean and standard deviation of the original data, were used to populate this new dataset. This approach aimed to test the algorithm's robustness and confirm that the peculiarities of the original dataset did not bias the results. In the SA algorithm, the weights applied were $w_c = 2$, $w_e = 1$, $w_c = 1$, and $w_{\text{inter}} = 0.75$. These were chosen to underscore the importance of the conscientiousness dimension while ensuring balanced representation across other dimensions.

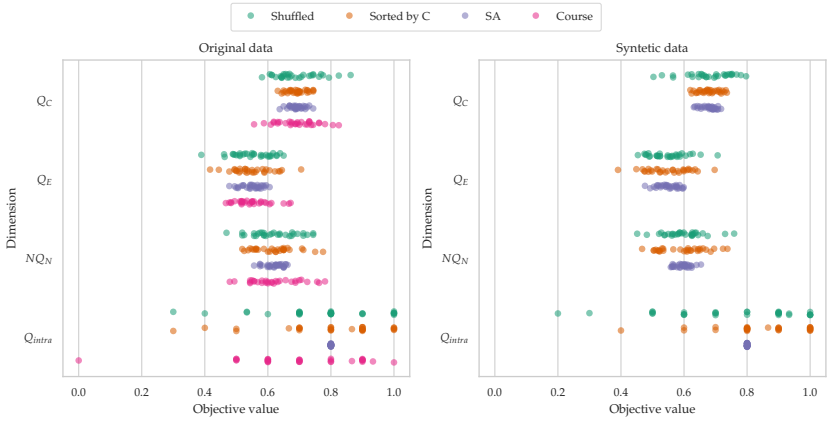


Figure 11. Distribution of scores before and after SA algorithm using 10,000 iterations, with weights of the objective function $w_c = 2$, $w_e = 1$, $w_c = 1$, and $w_{\text{inter}} = 0.75$

In the analysis of the original data, the *Sorted by conscientiousness* method yielded the lowest score, with $E(s) = 1.66$, followed by the *Shuffled* and *Course* methods, each scoring $E(s) = 1.70$. In contrast, the SA algorithm consistently outperformed these approaches, achieving an average score of $E(s) = 2.14$. Notably, the conscientiousness dimension, evaluated separately, matched the performance of the third method, with $E(s)_c = 0.67$. This indicates that the SA algorithm effectively sorts students based on conscientiousness and maintains a balanced distribution across the other dimensions.

To contextualize the computational efficiency of our algorithm, one can consider the total number of unique team configurations that would be examined in an exhaustive combinatorial search. Utilizing the formula $\binom{\text{number of students}}{\text{number of teams}}$, an exhaustive evaluation for selecting just one team of 5 from a pool of 150 participants would necessitate investigating 591,600,030 distinct combinations. In stark contrast, the SA algorithm substantially mitigates this computational demand. Specifically, it required searching only $710 \times 30 - 1740 = 19{,}560$ configurations to achieve the results presented in Figure 11.

6. Discussion

Our research focused on the Five-Factor Model of personality traits, hypothesizing that these data could positively affect team formation. Numerous studies support this idea [4,9–12,14,15], showing the impact of personality traits on team dynamics and performance. Our findings indicate that traits like conscientiousness and extraversion tend to improve team performance, while neuroticism tends to hinder it. Different studies highlight various traits. For example, [14] also found that extraversion and openness positively impact team performance but other traits had a lesser effect.

In contrast, ref. [11] pointed out the importance of conscientiousness but found other traits less significant. These differences might be due to the different study environments, participant groups, and research methods. As [12] notes, gender is important in team dynamics. However, many studies, ours included, mostly have participants of one gender, which limits the ability to consider gender differences fully.

Our study distinguished itself from others through our approach to data utilization. Beyond collecting psychometric data, we integrated information from Git repositories and team satisfaction surveys over an extended period. This integration provided a comprehensive view of team dynamics and performance. Utilizing this dataset, we developed a model capable of predicting team satisfaction with accuracy of 74% and precision of 69%. Despite the challenges posed by the limited size of our dataset and the complexities inherent in modelling human interactions, these performance metrics are noteworthy.

There were confounding factors not fully accounted for in our study, such as the technical skills of the students and their pre-existing relationships. Given that TF was largely random, it is plausible that some students had prior connections through earlier

courses or projects, potentially influencing team dynamics. Although our model has some limitations, its practical application demonstrated potential for success. When applied repeatedly within the TF algorithm, the predictive model effectively contributed to the formation of better-performing teams.

7. Conclusions

This study explored the impact of the Five-Factor Model of personality traits on team dynamics, analysing data from 157 third-year undergraduates formed into 31 teams. The results showed that teams typically perform better with more extroverted and conscientious but less neurotic members. Using an XGBoost model, we successfully predicted team satisfaction with 74% accuracy and 69% precision. We also introduced an innovative method for automatically forming teams based on the FFM. By applying a Simulated Annealing technique, we developed an efficient algorithm that effectively groups participants according to specific criteria. This method ensures a well-balanced distribution of personality traits among teams and enhances overall member satisfaction.

Our study's approach involved collecting a diverse range of variables about the participants, but the limitations primarily stem from the scope of the data collected. To enhance the model's precision, a broader dataset is essential. A single course was used, which limited the generalization of our findings. Expanding the range of variables to include factors like technical skills and gathering data from various courses with a more diverse participant group could mitigate gender biases and other disparities. Additionally, the study did not sufficiently focus on the impact of participants' technical abilities and the influence of varied roles within SE projects on team dynamics, since the project required all students to perform similar tasks.

Our recommendation for future research is to broaden the scope of the study by including a more comprehensive range of courses and a larger, more diverse group of participants. Additionally, integrating data on technical proficiencies and specific roles within teams could offer deeper insights into the nuances of team dynamics.

Author Contributions: Conceptualization, D.L. and J.V.; methodology, D.L.; software, J.V.; validation, J.V. and D.L.; formal analysis, J.V. and D.L.; investigation, J.V. and D.L.; data curation, J.V. and D.L.; writing—original draft preparation, J.V. and D.L.; writing—review and editing, D.L. and J.V.; visualization, J.V.; supervision, D.L.; project administration, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are not publicly available due to privacy restrictions. The participants in this research self-reported their answers regarding personality traits, and ensuring the confidentiality and privacy of their responses is of utmost importance. To protect the identity and sensitive information of the participants, we are unable to share the raw data.

Acknowledgments: Sincere gratitude goes to everyone who has contributed to completing this research, including the participating students and especially Marko Poženel and Aljaž Zrnec, for their dedicated assistance and support during the course Software Engineering.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TF	Team formation
SE	Software Engineering
MBTI	Myers–Briggs Type Indicator
FFM	Five-Factor Model

O	Openness
C	Conscientiousness
E	Extraversion
A	Agreeableness
N	Neuroticism
ICC	Intraclass Correlation Coefficient

References

1. Zainal, D.; Razali, R.; Mansor, Z. Team Formation for Agile Software Development: A Review. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2020**, *10*, 555. [CrossRef]
2. Budak, G.; Kara, I.; Ic, Y.; Kasimbeyli, R. New mathematical models for team formation of sports clubs before the match. *Cent. Eur. J. Oper. Res.* **2019**, *27*, 93–109. [CrossRef]
3. Sommerville, I. *Software Engineering*, 9th ed.; Addison-Wesley: Bruke, WA, USA, 2010.
4. Costa, A.; Ramos, F.; Perkusich, M.; Dantas, E.; Dilorenzo, E.; Chagas, F.; Meireles, A.; Albuquerque, D.; Silva, L.; Almeida, H.; et al. Team Formation in Software Engineering: A Systematic Mapping Study. *IEEE Access* **2020**, *8*, 145687–145712. [CrossRef]
5. Fiore, S.; Salas, E.; Cuevas, H.; Bowers, C. Distributed coordination space: Toward a theory of distributed team process and performance. *Theor. Issues Ergon. Sci.* **2003**, *4*, 340–364. [CrossRef]
6. Mathieu, J.E.; Rapp, T.L. Laying the foundation for successful team performance trajectories: The roles of team charters and performance strategies. *J. Appl. Psychol.* **2009**, *94*, 90–103. [CrossRef] [PubMed]
7. Burke, S. Is there a "Big Five" in Teamwork? *Small Group Res.* **2005**, *36*, 555–599.
8. Fiore, S.M.; Schooler, J.W. Process mapping and shared cognition: Teamwork and the development of shared problem models. In *Team Cognition: Understanding the Factors that Drive Process and Performance*; American Psychological Association: Washington, DC, USA, 2004.
9. Driskell, J.E.; Goodwin, G.F.; Salas, E.; O'Shea, P.G. What makes a good team player? Personality and team effectiveness. *Group Dyn. Theory Res. Pract.* **2006**, *10*, 249–271. [CrossRef]
10. Kichuk, S.L.; Wiesner, W.H. The big five personality factors and team performance: Implications for selecting successful product design teams. *J. Eng. Technol. Manag.* **1997**, *14*, 195–221. [CrossRef]
11. O'Neill, T.A.; Allen, N.J. Personality and the Prediction of Team Performance. *Eur. J. Personal.* **2011**, *25*, 31–42. [CrossRef]
12. Gilal, A.; Jaafar, J.; Omar, M.; Basri, S.; Izzatdin, A. Balancing the Personality of Programmer: Software Development Team Composition. *Malays. J. Comput. Sci.* **2016**, *29*, 145–155. [CrossRef]
13. Cruz, S.; Da Silva, F.; Monteiro, C.; Santos, C.; Dos Santos, M. Personality in software engineering: Preliminary findings from a systematic literature review. In Proceedings of the 15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011), Durham, UK, 11–12 April 2011; IET: Stevenage, UK, 2011; pp. 1–10. [CrossRef]
14. Soomro, A.B.; Salleh, N.; Nordin, A. How personality traits are interrelated with team climate and team performance in software engineering? A preliminary study. In Proceedings of the 2015 9th Malaysian Software Engineering Conference (MySEC), Kuala Lumpur, Malaysia, 16–17 December 2015; pp. 259–265. [CrossRef]
15. Rehman, M.; Mahmood, A.K.; Salleh, R.; Amin, A. Mapping job requirements of software engineers to Big Five Personality Traits. In Proceedings of the 2012 International Conference on Computer & Information Science (ICCIS), Kuala Lumpur, Malaysia, 12–14 August 2012; pp. 1115–1122. [CrossRef]
16. Scott, T.J.; Tichenor, L.H.; Bisland, R.B.; Cross, J.H. Team dynamics in student programming projects. *ACM Sigcse Bull.* **1994**, *26*, 111–115. [CrossRef]
17. Costa, P.; McCrae, R. The Five-Factor Model, Five-Factor Theory, and Interpersonal Psychology. In *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012; pp. 91–104. ISBN 9780470471609. [CrossRef]
18. John, O.P.; Srivastava, S. The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research*; Guilford Press: New York, NY, USA, 1999.
19. Ashton, M.; Lee, K.; Goldberg, L.; de Vries, R. Higher Order Factors of Personality: Do They Exist? *Personal. Soc. Psychol. Rev.* **2009**, *13*, 79–91. [CrossRef] [PubMed]
20. Jia, J.; Zhang, P.; Zhang, R. A comparative study of three personality assessment models in software engineering field. In Proceedings of the 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 23–25 September 2015; pp. 7–10. [CrossRef]
21. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
22. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
23. Luke, S. *Essentials of Metaheuristics*, 2nd ed.; Lulu: Morrisville, NC, USA, 2013.
24. Hancock, S.A.; Hill, A.J. The effect of teammate personality on team production. *Labour Econ.* **2022**, *78*, 102248. [CrossRef]

25. Linden, D.v.d.; Nijenhuis, J.t.; Bakker, A.B. The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *J. Res. Personal.* **2010**, *44*, 315–327. [CrossRef]
26. *ICERI 2015: 8th International Conference of Education Research and Innovation, Seville (Spain), 16–18 November 2015: Proceedings*; Iated Academy: Casablanca, Morocco, 2015.
27. Liljequist, D.; Elfving, B.; Roaldsen, K.S. Intraclass correlation: A discussion and demonstration of basic features. *PLoS ONE* **2019**, *14*, e0219854. [CrossRef] [PubMed]
28. Cicchetti, D. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology. *Psychol. Assess.* **1994**, *6*, 284–290. [CrossRef]
29. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]