Unveiling Career Pathways: Success and Challenges of Bangladeshi Women in Computer Science Through Machine Learning

Md. Azim Howlader[1], Fahad Bin Aziz[2], Mahbuba Yesmin Turaba[3*], Tarikuzzaman Emon[4], Moskura Hoque[5], Israt Jahan[6], Ahmed Saif Reza[7], Rayhan Chowdhury[8]

[1,6,8] Department of CSE, World University of Bangladesh, Uttara, Dhaka, Bangladesh
[2,5]Central University of Science and Technology, Mirpur, Dhaka, Bangladesh
[3]C3Bit, Dhaka, Bangladesh
[4]Department of CSE, Stamford University Bangladesh, Siddeswari Road, Dhaka, Bangladesh
[7]Medico Bio Limited, KalaBagan, Dhaka, Bangladesh

[*]**Email:** turaba@c3bit.com

Abstract

Although the number of working computing women is steadily increasing in Bangladesh, it is a ray of hope that the gender gap is reducing day by day among final year students to higher-level job holders. This research aims to forecast how women in Bangladesh perceive and respond to pursuing careers in Computer Science. Primary data is collected by surveying women's experience that incorporates various open-ended and closed-ended questions and thus developed a dataset from 501 respondents, whereas respondents' age group were 19 to 60 years, and the majority were working in private sector jobs. A statistical tool Pearson's chi-square test is implemented to correlate between variables and thus different machine learning approaches, including Random Forest (which achieved a topmost accuracy of 85.00%), Decision Tree, XGBoost, Logistic Regression and K-Nearest Neighbors were implemented. It has explored the position, success, and obstacles of women in their place in Computer Science in Bangladesh, and one of the most delightful revelations is the borderline association of unequal pay. Notably, over 66% of the respondents reported that they do not encounter gender-based discrimination in their workplaces in terms of career advancement within various sectors of computer science.

Keywords

Career Success and Challenges, Pearson's chi-square test, Machine Learning Models, Gender Gap, Mixed-mode research

Introduction

Bangladesh is developing rapidly in Asia and has 500,000 active out of 273,000 registered freelancers, and places second position according to a study by the Daira (Bangladesh Institute of

Submission: 30 August 2025; **Acceptance:** 23 September 2025; **Available online:** September 2025

Governance and Development) (Arman et al. 2025). It is evident that we should focus on women, their socio-economic empowerment, and the value they bring to the workplace to decrease poverty and boost economic development (Wei et al. 2021). Women face challenges due to not having good technical tools, a lack of support from family, low confidence in programming, and inequality in the workplace (Guzmán et al. 2024). The elements, including education, occupation and circumstances, could impact their success rate. Women are sometimes underrepresented because their lack of experience and resources prevents them from pursuing professions in technology. To increase women's representation from undergraduate to industry positions, the whole spectrum, including academics, policymakers, and society, is working heart and soul (Jaccheri et al. 2020). Moreover, society has given women certain roles to play, and these roles often get in the way of their growth (Fitong et al. 2023).

In this research, we have explored multiple machine learning models to predict success in CS based on wages, positions, knowledge and confidence in computer programming languages. The main contribution of our research is summarized below-

- Our primary data is collected by surveying women's experiences and developing a dataset from 501 respondents. Our respondents' age group were 19 to 60 years, and the majority of them were private job holders.
- Statistical analysis Chi-Square test depicts significant association between Age, Occupation, Proficient in programming, CGPA, Projects, Challenges, Team leader, Teammate and Extra hour.
- To bridge the gender gaps, machine learning techniques are utilized to investigate how successful women with a computer science background are in their careers and to encourage those who are still afraid to focus on this sector.
- Five essential machine learning models, such as Decision Tree, Random Forest, XGBoost, Logistic Regression, K-Nearest Neighbors (Ibarra et al. 2024) are explored to predict the success and growth of women and Random Forest is the topmost performing model in terms of accuracy, precision, recall and F1 score.
- Success and challenges faced by women in the CS industry are examined, as well as the strategies they can employ to overcome these obstacles. Thus, the success rate prediction is based on age, family support, position at work, gender discrimination and so on.

Methodology

Data collection, Data preprocessing, Statistical analysis Chi-Square Test, Machine learning models, and women's career success prediction make up the whole process for career progress assessment. Thus, the methodology is depicted in Figure 1.

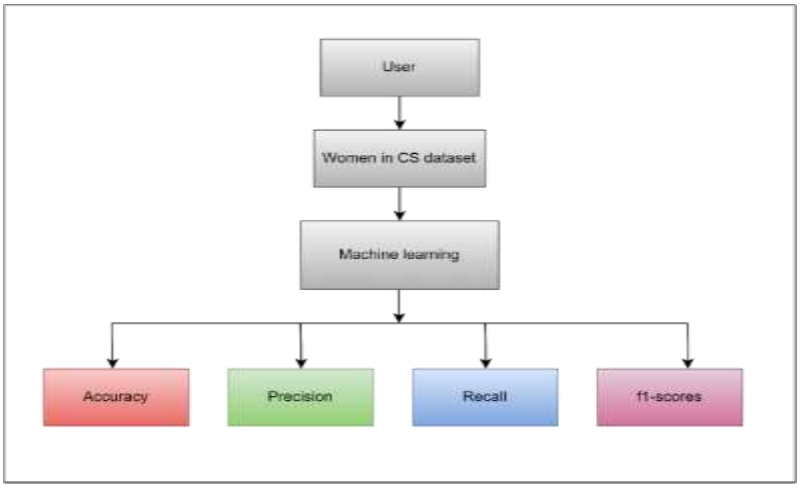

Figure 1: This proposed method demonstrates building dataset, machine learning models, and evaluation metrics.

Data Collection

In Bangladesh's CS job sector, women's participation is less than that of men. Data is collected using a mode response from women with a computer science background in Bangladesh. We incorporated open-ended and closed-ended questions, and we reached out to 20% respondents in person by researcher administered questionnaire, while 80% respondents filled out the online questionnaire form.

Participants

The survey aimed to include over 500 participants, and we successfully collected data from 501 respondents, which contributes to enhancing the performance of our machine learning model. Data collection commenced on 15 July 2024 and continued till June 2025. The survey was administered using three distinct methods. The first way was the conventional one, in which participants completed a survey form that we provided to them. In the second way, the same questionnaire pattern in Google Form was developed to collect this data. Another way is Focused Group Discussion (FGD) for in-depth insights from respondents. Moreover, to collect data, we visited different tech companies and reached women from various positions and have also collected responses through Social Media.

Dataset Description

There are a total of 17 columns in this dataset. A few questions had three categories for answers, which are Yes, No and Moderate. Some of the questions contain nominal variables, either “Positive”, “Negative”, "Neutral” or ordinal categories "Entry", "Mid" and "Higher". In this dataset, we found that most of the respondents are working in private jobs, their age range is between 19-30, and most of them are at the entry level. A number of respondents are teachers, freelancers and job seekers after completing an undergraduate level. Most of them reported substantial family support and flexibility to work extra hours if necessary. Regarding programming skills, the majority indicates a level of “Moderate”. Additionally, approximately 90% of women claim that they could work as good teammates as well as team leaders. It is worth mentioning here

that, although gender discrimination in the workplace is relatively low, most responses indicate that women are not paid as expected in the workplace.

Sample Size Determination

To calculate the sample size for an unknown population for research purposes, 95% confidence level and Z-score: 1.96 with confidence Interval: 0.05 and a Standard Deviation: 0.5 (Safe choice since the figure is unknown), we could calculate the Sample size as follows,

$$n = \frac{\left(Z^2 * p * (1 - p)\right)}{E^2}$$

Whereas n denotes the sample size, Z represents the Z-score corresponding to the desired confidence level, p is the estimated proportion of the population with the characteristic of interest, and E indicates the error margin (desired level of precision). (Mukti et al. 2025) Thus, we get n=384.16.

Data Preprocessing

At first, missing entries are examined. Then, missing values were filled with the mode (most frequently occurring category) for categorical variables, and for numerical variables, missing values were imputed using the mean of the respective variable. Therefore, transformation has been applied, and our categorical data is converted into numerical format by the Label Encoder technique to enhance the dataset's reliability, and for nominal categorical variables, One-Hot Encoding is employed (Poslavskaya et al. 2023). The overall steps for data preprocessing are outlined in Figure 2.

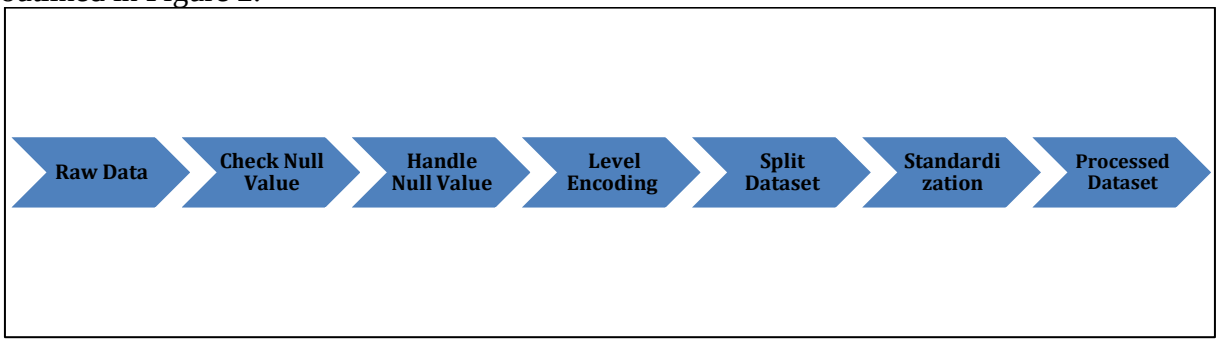

Figure 2: Steps for data preprocessing include null value check and handling, label encoding, and standardization

Chi-Square test

The characteristics of the participants have been obtained by employing descriptive statistics, specifically frequencies and percentages, in order to analyze categorical data.
Hypothesis:
H0: This variable is no association between a variable and the Career success of women
H1: This variable is an association between a variable and the Career success of women. (Aslam et al. 2023)

Thus, the Hypothesis test is visualized in Table 1, where we have employed all features. Variables such as Honors, Programming, Discriminate, Remuneration, Perfectionist, Thesis projects, and Marital status are not significantly correlated, while others are significantly correlated.

Table1: A Pearson's chi-square test depicts the association between the total 17 variables and the career success of women

Variables	Chi-square value	Degrees of freedom	P value
Age	4.366	2	0.0112
Occupation	45.900	14	0.00002
Honors	1.068	2	0.58602
Proficient	33.173	4	0.000001
CGPA	23.236	2	0.000009
Programming	4.736	4	0.31542
Discriminate	4.642	2	0.0981
Projects	6.452	2	0.03971
Remuneration	5.916	2	0.05190
Challenges	12.562	2	0.00187
Perfectionist	2.586	2	0.27433
Team leader	12.450	2	0.00197
Teammate	13.447	4	0.00928
Extra hour	19.631	4	0.00590
Thesis projects	6.187	6	0.40253
Marital status	1.164	2	0.55851
Family support	1.475	2	0.47813

Machine learning models

The inclusion of a variety of machine learning models, such as Decision Tree, Random Forest, XGBoost, Logistic Regression, and K-Nearest Neighbors to comprehensively analyze various factors affecting women's success.

Result and discussion

Experimentation involves training models on 80% of the dataset and testing on the remaining 20%. The inclusion of a range of evaluation metrics, including accuracy, precision, recall, and the F1 score, ensures a thorough assessment of the model's performance (Turaba et al. 2022). After training the model, the Random Forest technique achieves the topmost accuracy with 85.00% in the complex dataset and leads to reduced overfitting. Figure 3 depicts the overall result for all machine learning models.

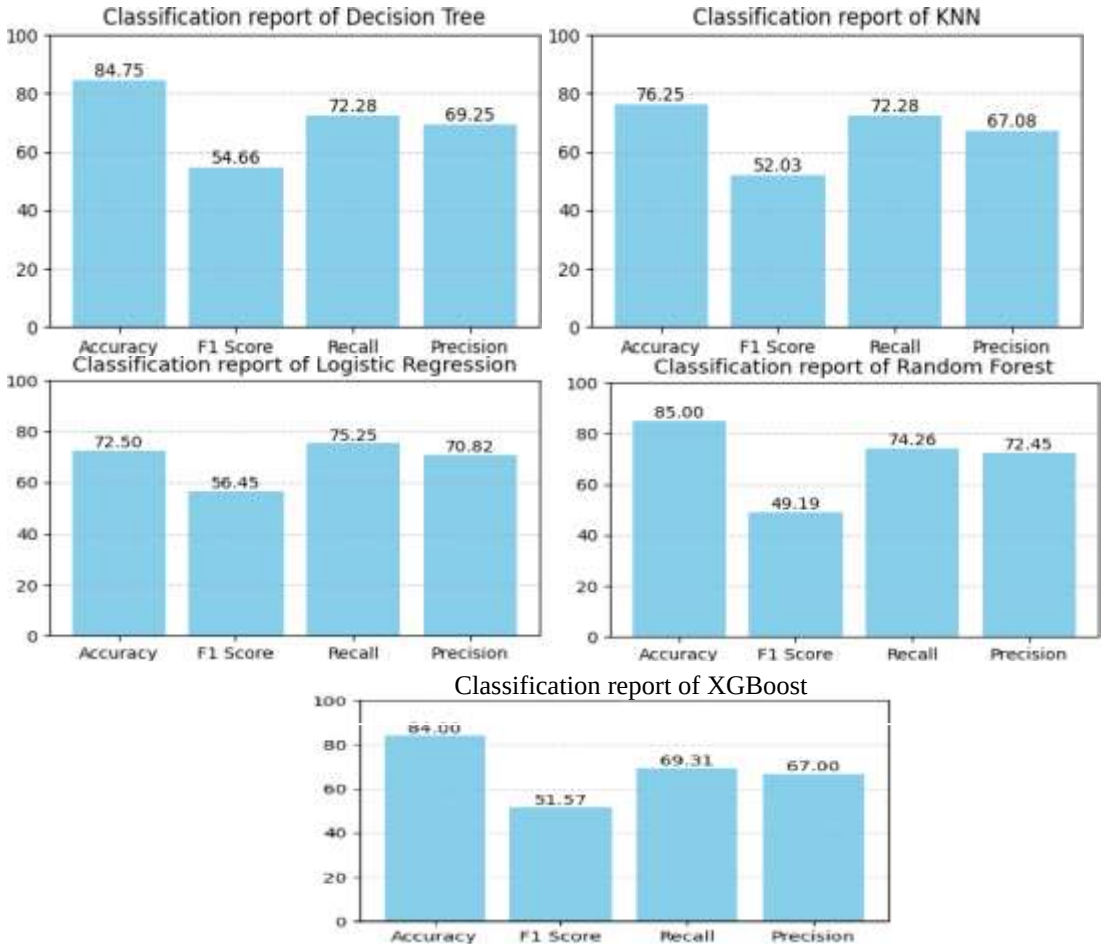

Figure 3: The performance of different models was evaluated in terms of accuracy, precision, recall, and F1-score.

It is noteworthy that throughout the hiring process, employers recommended candidates with proficiency in programming, higher CGPA, challenge taker and teammate to the team. Therefore, some women get financial assistance from their families, while others are under intense pressure to perform well so that they might climb the corporate ladder with their skill sets.

Conclusion

In wrapping up our research, we have gained valuable insights into women in Bangladesh's CS sector through conversations with over 500 women, typically between the ages of 19 and 60 years, whereas women remain underrepresented in the CS industry compared to their male counterparts. Regarding technical proficiency, while most respondents' express confidence in their programming abilities, a substantial portion still harbor feelings of inadequacy and believe they could further enhance their skills. Additionally, it is disheartening to note that 44% women continue to encounter workplace discrimination, although its prevalence appears to be on the

decline. It is imperative to foster an environment of inclusivity, gender equality and respect to ensure that all individuals feel valued and treated fairly in the workplace, despite their qualifications and contributions, whereas 40.1% women feel they are fairly remunerated. Through rigorous statistical analysis Chi-Square test, and leveraging advanced techniques such as Random Forest, we successfully developed predictive models that accurately classify these job levels, highlighting the potential for data-driven solutions to address challenge-taking mentality, good team player and leader in the workplace. Future extension could be collecting more data to build a larger dataset, focusing on mental well-being and implementing more machine learning and deep learning algorithms to optimize results to identify how to improve the development of women in the private sector, government jobs and teaching profession in CS.

Acknowledgements

There is no grant or funding bodies to be acknowledged for preparing this paper. The authors acknowledge with appreciation the valuable assistance of the research team from Stamford University, Bangladesh, in facilitating data collection. Their support and scholarly engagement also provided significant inspiration for undertaking this research.

References

Arman, M. R. (2025). Transforming the workforce: Enhancing employability of women and madrasa graduates through online freelancing. Daira. https://dairabd.org/transforming-the-workforce-enhancing-employability-of-women-and-madrasa-graduates-through-online-freelancing/

Aslam, M., & Smarandache, F. (2023). Chi-square test for imprecise data in consistency table. Frontiers in Applied Mathematics and Statistics, 9, 1279638. https://doi.org/10.3389/fams.2023.1279638

Fitong Ketchiwou, G., & Dzansi, L. W. (2023). Examining the impact of gender discriminatory practices on women's development and progression at work. *Businesses*, *3*(2). https://doi.org/10.3390/businesses3020022

Guzmán, E., Fischer, R. A. L., & Kok, J. (2024). Mind the gap: gender, micro-inequities and barriers in software development. *Empirical Software Engineering*, *29*(1), 17. https://doi.org/10.1007/s10664-023-10379-8

Ibarra-Vazquez, G., Ramírez-Montoya, M. S., & Terashima, H. (2024). Gender prediction based on University students' complex thinking competency: An analysis from machine learning approaches. *Education and Information Technologies*, *29*(3), 2721-2739. https://doi.org/10.1007/s10639-023-11831-4

Jaccheri, L., Pereira, C., & Fast, S. (2020). Gender issues in computer science: Lessons learnt and reflections for the future. In Proceedings of the 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2020) (pp. 9–16). http://dx.doi.org/10.48550/arXiv.2102.00188

Mukti, B. H. (2025). Sample size determination: Principles and applications for health research. Health Sciences International Journal, 3(1), 127–143. https://doi.org/10.71357/hsij.v3i1.63

Poslavskaya, E., & Korolev, A. (2023). Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding? https://doi.org/10.48550/arXiv.2312.16930

Turaba, M. Y., Hasan, M., Khan, N. I., & Rahman, H. A. (2022, October). Fraud detection during financial transactions using machine learning and deep learning techniques. In Proceedings of the 2022 International Conference on Computing, Communication, and Intelligent Systems (CCCI) (pp. 1–8). IEEE. https://doi.org/10.1109/CCCI55352.2022.9926503

Wei, W., Sarker, T., Żukiewicz-Sobczak, W., Roy, R., Alam, G. M., Rabbany, M. G., ... & Aziz, N. (2021). The influence of women's empowerment on poverty reduction in the rural areas of Bangladesh: Focus on health, education and living standard. *International journal of environmental research and public health*, *18*(13), 6909. https://doi.org/10.3390/ijerph18136909