JOURNAL OF
EMERGING INVESTIGATORS

Article

Understanding the impossibility of machine learning fairness with data examples

Keshav Pillutla [1], Kevin Fry [2]
[1] Mountain View High School, Mountain View, California
[2] Stanford University, Stanford, California

SUMMARY

Machine learning has become increasingly prevalent in the world as a result of a combination of factors, notably computing breakthroughs and increased data availability. The three most common criteria of fairness in machine learning, despite their common sense and moral appeal, are often mutually exclusive. This frequently presents a challenge and the need to prioritize one criterion over the others. Specifically, the paper highlights the general overview of a machine learning model and presents an example of its application in the legal field that explores the existing biases in models today. We then delve into the three main criteria of machine learning fairness: independence (fairness based on outcomes being unrelated to different characteristics), separation (fairness based on equal error rates across groups), and sufficiency (fairness based on predictions being equally reliable for all groups). In the present experiment, we aimed to investigate the mutual exclusivity of these three machine-learning fairness criteria. We hypothesized that the fairness criteria being evaluated cannot all three be satisfied simultaneously, leading to a machine learning model that remains unfair. The results demonstrated that no threshold in the model simultaneously satisfied independence, separation, and sufficiency, highlighting the limitations of machine learning models that pose various issues across different sectors.

INTRODUCTION

A machine learning model takes information and data from the past and uses it to predict trends and future outcomes of a certain event. Some researchers, such as those from ProPublica who have explored machine learning fairness in depth, have theorized that discrimination and demographic disparities have influenced these predictions (1). The societal aspect of machine learning raises the question of the impossibility of fairness, due to the inevitable bias of humans associated with artificial intelligence (AI). Machine learning fairness refers to the concept of a perfect machine learning model that can predict trends without bias (2). This becomes extremely important in fields such as law, healthcare, finance, and employment, where algorithmic decisions can significantly affect people's lives (1). The three most common fairness criteria are independence, separation, and sufficiency (3). Independence is the idea that the probability that the model will predict an event to happen is independent of changes to a sensitive characteristic (3). A sensitive characteristic is a defining characteristic in a human being that could cause bias in a model (4). This means the probabilities (or error rates) will be the same even if one changes that sensitive characteristic (3). For example, if the model was trying to predict whether someone would get accepted into college, the probability that the model predicts they would get in should not change if this person's race changes. Separation is similar, where an event's predictive probability must be the same even if the sensitive characteristic changes, given that the event has already happened (3). This means that the sensitive characteristic changing will not affect the true positive or false positive rates in a model, as these rates are computed by calculating the predictive probabilities given an event has or has not happened (3). If, for example, we were to take a male and a female diagnosed with COVID-19, separation is the concept that the probability that they are and are not predicted to have that disease must be the same. Because of this, the true and false positive rates and true and false negative rates must be equal for males and females. Finally, sufficiency is the notion that the likelihood of something happening is the same when the sensitive characteristic changes, given that the model predicted the event would happen regardless of the characteristic (3). This means that both the positive predictive value and negative predictive value are equal across different groups or sensitive characteristics, as these two values are calculated using the proportion of correct positive and negative predictions (3). Let us return to the example of college admission and use disability as our sensitive characteristic. The notion of sufficiency states that two people, one without disability and one with disability, must have the same probability of getting into college, given that they were predicted to be admitted. Unlike separation, this equalizes the positive and negative predictive values, the false omission, and the false discovery rate. Prior research has both mathematically and theoretically proven that satisfying these three criteria simultaneously is often impossible in real-world systems, meaning that a perfect machine learning model is something that we cannot achieve (5).

When discussing machine learning fairness, the concept of sensitive variables, or characteristics, generally comes up. It may be considered unfair when a computer decision is based on these variables. Gender, ethnicity, sexual orientation, and disability are examples of sensitive variables (4). This matters because when computers use these sensitive characteristics in machine learning, they introduce societal inequalities into the system, causing one race or gender, for example, to be judged more harshly than another (1). Rather than implementing discriminatory beliefs into a machine learning model, a fair model would address these biases and ensure that its decisions align with the fundamental human values of equality. However, this type of model is only hypothetical,

https://doi.org/10.59720/25-166

as achieving this goal is often mathematically impossible (5). Simply ignoring these biases is not a viable solution. A study has shown that it is challenging to ignore race, for example, as a sensitive characteristic, as variables that have a correlation to race may still cause racial disparities in the data, and therefore the model (6). Even if these characteristics were ignored, the bias lies in the outcomes, meaning that the patterns that stem from racial inequality still remain in the model.

The issue of machine learning fairness emerged through the controversy with the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) AI model, initially discovered by the investigative journalism non-profit, ProPublica (1). COMPAS is an AI model that predicts the likelihood of recommitting a crime based on many factors associated with the criminal being assessed (1). These assessments influenced decisions regarding a defendant's release, bond amount, and sentence severity (1). Race, among many other factors, was a sensitive characteristic causing bias in the model (1). For example, people of color were more often predicted to commit a crime again compared to white criminals, even if their criminal records said otherwise as shown by Propublica's analysis of COMPAS data from Broward County, Florida (1). The ProPublica article covers many instances of people of color being charged with the label that they were at elevated risk of committing another crime in the future. It described how black defendants were twice as likely to be wrongfully labeled as high-risk than white defendants (1). Although factors such as criminal history, age, and gender could potentially account for this difference, further statistical analysis revealed that the racial disparities persisted even after controlling for these other variables, significantly influencing risk assessment outcomes (1). The system was intended to reduce human bias in the criminal justice process but instead highlighted how algorithmic bias still exists and preserves existing inequalities. This connects back to the fairness criteria of independence, separation, and sufficiency, as the model violates them. This issue brought the uncertainty centered around machine learning fairness to the public eye, raising the question of whether it was possible to truly create a perfect and fair machine learning model. An example of the argument for its possibility is a paper trying to disprove the theory of mutual exclusivity of the fairness criteria (7). This paper predominantly looks at other studies on this subject and gives us reasoning as to why these studies either had limitations or false information (7). It looks at the different cases listed in this article from another, much broader perspective, to give us an understanding of how COMPAS did not have bias and was fairly accurate when judging both black and white defendants (7). Although this may be true, COMPAS is one specific case of a much larger problem, meaning the question of machine learning fairness and whether it is possible to achieve it still stands.

The empirical data resulting from the COMPAS case, along with the general information mentioned in this section has caused widespread doubt about the future of AI in our world, and more specifically, how it could adapt to society's biases against certain groups of people. We hypothesized that the fairness criteria under consideration are unable to be satisfied simultaneously, resulting in a biased and unfair machine learning model. This study seeks to investigate whether machine learning models can make decisions without bias playing a role. The mathematical aspect of this idea is beyond the scope of this paper, but another way we can verify this is through using a real-world dataset and calibrating our model to try and simultaneously satisfy all three fairness criteria considering only binary outcomes. We explored this aim by using a public dataset on student success that includes data on 3630 students and their status, including characteristics such as gender, marital status, financial status, etc., and we also used a program written in Python (8). Then, we examined this case study to understand how the model cannot satisfy all three fairness criteria simultaneously due to correlations between the sensitive variables and other predictive features in the dataset, which embed biases into the data itself. To this end, this study explored a model's fairness criteria and demonstrated its mutual exclusivity, proving that a fair machine learning model is often impossible to achieve.

RESULTS

The aim of this study was to test the hypothesis of the mutual exclusivity of the independence, separation, and sufficiency. We evaluated whether a logistic regression model could simultaneously satisfy these three most common

Criterion being satisfied	Threshold(s) Where Criterion is Satisfied	True Positive Rates (males, females)	False Positive Rates (males, females)
Independence	49.5%, 74.9%	77.1%, 67.8%	46.6%, 47.9%
Separation	2.6%	96.6%, 96.6%	69.4%, 69.4%
Sufficiency	61.4%, 80.2%	12.7%, 32.5%	4.6%, 30.6%

Table 1. Thresholds for satisfaction of independence, separation, and sufficiency in the logistic regression model. Describes the thresholds, true positive, and false positive rates where each fairness criterion (independence, separation, and sufficiency) is satisfied. Independence is defined as the difference in proportion of students dropping out. Separation is defined as the difference in false positive rates and true positive rates, respectively, for males and females. Sufficiency is defined as the difference in positive predictive value for males and females. For independence and sufficiency, two values indicate two thresholds in which these criteria were satisfied.

Threshold(s)	Independence	Separation	Sufficiency
49.5%, 74.9%	Satisfied	1.3%, 9.3%	16.7%
2.6%	18%	Satisfied	13.9%
61.4%, 80.2%	23.5%	26%, 19.8%	Satisfied

Table 2. Fairness criteria tradeoff error chart. Error chart displaying how much error is in the other fairness criteria when one criterion is satisfied (what threshold or thresholds independence, separation, or sufficiency are taking place). Independence is defined as the difference in proportion of students dropping out. Separation is defined as the difference in false positive rates and true positive rates, respectively, for males and females. Sufficiency is defined as the difference in positive predictive value for males and females. For independence and sufficiency, two values indicate two thresholds in which these criteria were satisfied.

https://doi.org/10.59720/25-166

fairness criteria when predicting student graduation status. The results summarize how model performance varied with respect to each fairness criterion across prediction thresholds.

A confusion matrix is a grid that evaluates a classification model by comparing its predicted labels against its actual labels. In this study, we analyzed a confusion matrix to investigate the true and false positive rates needed to assess the fairness criteria. These values were eventually plotted on a Receiver Operating Characteristic (ROC) curve, which plots the relationship between false positive and true positive rates across various thresholds. The confusion matrix demonstrates that there was no point on the ROC curve where independence, separation, and sufficiency were satisfied (**Table 1**). This indicates that there is no threshold we could have used for this machine learning model that would have allowed it to be fair and unbiased. The matrix shows that independence was satisfied at the thresholds of 74.9% and 49.5% when predicting that approximately 40% of the students would drop out. The error table displays how at these thresholds, the false positive rates were only 1.3% apart (**Table 2**). Still, the true positive rates differed, with a difference of around 9%, showing an error in separation, which is satisfied when the true and false positive rates are equal. Sufficiency at this threshold had an error of 16.7%, which is the difference in positive predictive values between males and females. The matrix also shows that separation was satisfied at around the threshold of 2%. When satisfied, the error table shows that there was an 18% difference in the proportion of students dropping out and a 13.9% difference in the positive predictive values, demonstrating how independence and sufficiency were not satisfied. The sufficiency criterion was met at the thresholds of 61.4% and 80.2%, as shown by the confusion matrix. When sufficiency was satisfied, the false and true positive rates differed, with a difference of 26% and 19.8%, respectively, showing a substantial violation of separation, as displayed in the error table. The proportion of students dropping out also differed by 23.5%, which shows a violation of independence.

Figure 1. Heatmap displaying fairness criteria trade-offs. Rows indicate which criterion is satisfied; columns show the violation magnitude (%) of each criterion. Diagonal cells show that each criterion has no error in satisfying itself. Off-diagonal cells show mutual exclusivity: when one criterion is satisfied, others are violated. For separation, two values represent differences in false and true positive rates, respectively.

https://doi.org/10.59720/25-166

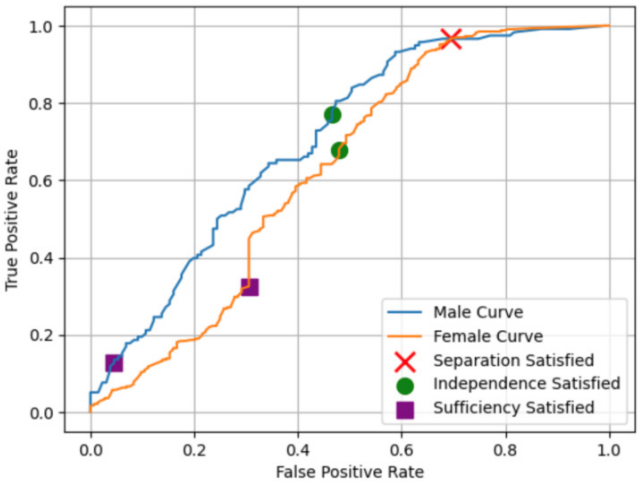

Figure 2. True positive and false positive rates where independence, separation, and sufficiency are satisfied for males and females. Receiver operating characteristic (ROC) curves for the model fit onto the student performance data. Green dots are where independence is satisfied; red 'x' is where separation is satisfied, and purple squares are where sufficiency is satisfied.

The percentage errors shown in the error table demonstrate the trade-offs presented when creating a machine learning model. In this case, satisfying independence may be a priority. This is based on the minimal differences between false and true positive rates when independence is satisfied, along with positive predictive values. However, satisfying any of these criteria depends on what a person considers the most important. The heat map also shows these errors, visually emphasizing how much the other criteria are affected when one is satisfied (**Figure 1**). The multiple areas that are shown to have high violation magnitude demonstrate the tradeoffs that are presented when one criterion is satisfied rather than others, which, in turn, shows how difficult it is to identify the best criterion to prioritize when creating a machine learning model.

The ROC curve visually displays the thresholds at which the three criteria are satisfied, with the "x" marking satisfaction of separation, the dot marking independence, and the square marking sufficiency (**Figure 2**). The model and the ROC curve indicate that it is impossible to audit our model to make it fair amongst all genders using solely these criteria.

DISCUSSION

The results of this study demonstrated that no prediction threshold simultaneously satisfied independence, separation, and sufficiency when running the logistic regression model. Instead, different thresholds satisfied different criteria, illustrating the trade-offs that must be considered when creating a fair machine learning model using real world data.

The calibration of the machine learning model used in this study demonstrated that no point on the ROC curve satisfied all three fairness criteria, exhibiting the trade-off that occurs when creating a machine learning model. At certain thresholds, we can only meet one of these fairness criteria, implying that when fitting a model to a dataset, one must consider which fairness criterion to satisfy, whether it be independence, separation, or sufficiency. Understanding these trade-offs is essential, as prioritizing one fairness criterion over another can lead to different outcomes in important real-world applications such as healthcare, hiring, or criminal justice, where misclassification can directly affect an individual's life. This study highlights the general issue of machine learning fairness. The implications of these results are extensive because they can impact the future of AI-assisted decision-making. The biases in machine learning models are potentially rooted in the fact that human bias can never fully be extinguished, as humans develop AI models. These doubts about bias in models relate to multiple studies based on the recurring issue of machine learning fairness in this world.

The results also demonstrated that the extent of violation differed depending on which fairness criterion was satisfied. When independence was satisfied, the model showed relatively small differences in false positive rates and positive predictive values, but exposed disparities in true positive rates, indicating a violation of separation. In contrast, satisfying sufficiency produced larger differences in both false positive and true positive rates, suggesting that separation was more heavily affected under the conditions of sufficiency being satisfied. Separation satisfaction also resulted in considerable differences in dropout proportions and predictive values, demonstrating violations of both independence and sufficiency. These findings suggest that certain fairness criteria may produce more substantial differences in other measures depending on which criterion is prioritized when creating the machine learning model.

These findings also demonstrate how machine learning fairness is not just a mathematical problem, but becomes a practical issue when models are applied to real-world situations. The different fairness criteria, by definition, prioritize different definitions of equity, meaning that the selection of one criterion over the other may influence how errors are distributed across certain groups. As a result, determining which fairness criterion to prioritize depends on the context in which the model is used, as the potential consequences associated with incorrect predictions must be considered. Therefore, because of the mutual exclusivity of these fairness criteria, it is vital to understand the context in which the model is being created so different forms of equity can be used to make important decisions.

This study also has several limitations. The analysis described in this study focused specifically on the three most commonly discussed fairness criteria. However, although these criteria primarily define machine learning fairness, they do not encompass every possible definition of fairness within machine learning (9). Additionally, this study examined fairness using a logistic regression model and a single dataset, meaning that the results may vary when applied to different datasets or machine learning approaches. It also exclusively focused on gender as a sensitive characteristic, whereas other characteristics such as race and economic status could produce different results. Further research may consider different approaches by creating multiple models to provide more conclusive results into the issue of machine learning fairness.

Another study of these three fairness criteria demonstrated that their mutual exclusivity can be disproven by mathematics

https://doi.org/10.59720/25-166

(5). Accordingly, the study provides a detailed analysis of each criterion and then demonstrates that these cannot be satisfied simultaneously, using mathematics and logic (5). It gives the trade-offs as a result of the mutual exclusivity of these criteria, along with finding constrained, special, hypothetical cases where they could be satisfied at the same time (5). Previous literature regarding these fairness criteria also details their mutual exclusivity from a logical standpoint (2). After introducing the concept of machine learning, it presents ideas about the historical aspects of society, like discrimination laws, along with multiple case studies to comment on how different notions of fairness can lead to conflicting outcomes in machine learning systems (2). These prior works introduced various methods of proving the mutual exclusivity of the fairness criteria. This further solidifies the claim that machine learning fairness is often impossible to achieve.

This study examined the trade-offs associated with machine learning fairness criteria and demonstrated their mutual exclusivity within a model. By auditing this model, through changing the threshold that allows the model to output the graduating status, we discovered that different thresholds satisfy different fairness criteria. However, the ROC curve demonstrates no overlap between these thresholds, implying that no one threshold can fulfill all three fairness criteria.

This study provides an empirical example of the recurring issue in the real world. As AI continues to grow in popularity, and as we continue to integrate it with many aspects of our lives, the questions surrounding its fairness in models become more prevalent. The controversy around machine learning fairness has caused a major uproar in society because of the impactful repercussions it can have, especially when it comes to discrimination (1). This is because biased models are influenced by existing prejudices, resulting in societal inequalities (10). An individual's race, gender, or sexual orientation should not define the outcome of a certain situation, which is what these three common fairness criteria are centered around. However, satisfying all three of these can only happen in theoretical cases. Although it is possible to satisfy these criteria under ideal conditions, real-world data contains too much difference in group distributions, the distribution in characteristics, scores, or features within a specific demographic group, and outcome base rates, the actual, real-world percentages of a specific outcome within a group (11). As a result, all three fairness criteria cannot generally hold true simultaneously, which is why achieving them together is considered impossible (11). This paper and many other studies preceding it demonstrated the phenomenon centered around the impossibility of machine learning fairness and the mutual exclusivity of independence, separation, and sufficiency.

MATERIALS AND METHODS

This section overviews the steps taken to calculate each fairness criterion, using code written on the platform Google Colab, which is available on GitHub (**Appendix**).

Fairness Criteria Definitions

The fairness criteria evaluated in this study were independence, separation, and sufficiency, and it was centered on satisfying these three simultaneously for predicting a student's status. The mathematical formula for independence can be written as follows:

$$P(r \mid R = r \mid A = a) = P(R = A = b) \quad \text{(Equation 1)}$$

Where R is set to the predicted outcome of r and A is a sensitive characteristic, which is the changing variable, being either a or b.

The mathematical formula for separation can be written as follows:

$$P(R = r \mid Y = q, A = a) = P(R = r \mid Y = q, A = b) \quad \text{(Equation 2)}$$

In this case, Y is set to the true outcome of q.

The mathematical formula for sufficiency can be written as follows:

$$P(Y = q \mid R = r, A = a) = P(Y = q \mid R = r, A = b) \quad \text{(Equation 3)}$$

Data Source and Processing

A dataset from Portugal was used that contained 3630 students who were studying for a diverse set of undergraduate degrees (8). This dataset focused on the students who dropped out or graduated, based on several factors, or sensitive characteristics, with this notebook focusing on gender. First, the main dataset was stratified by gender. Many variables were present in the dataset, such as nationality, marital status, parental occupations, economic status, etc. However, gender was used as the sensitive characteristic in this experiment to allow for a clearer distinction between the results for males and females.

The data was divided into training and testing data, with an 80/20 split (i.e., 80% of the data was used for training, and 20% was used for testing). Then, a generalized linear model was fitted onto the training data using logistic regression, a classification method in machine learning. After this, the full dataset was stratified by gender. A logistic regression was also performed on these two datasets. This means that based on the independent variables in the dataset, the probability of the situation, in this case graduating, was estimated through machine learning. Using these probability values, the three fairness criteria were computed. A logistic regression model was run to test these three fairness criteria because it provides understandable probability outputs that reveal relationships between predictors and protected attributes. This allows for easy analysis of these probabilities to find out the tradeoffs between the different fairness criteria. To run this logistic regression, the statsmodels module was used to create a generalized linear model. This meant that the model was not trained iteratively over a specific number of epochs, but rather using the Iteratively Reweighted Least Squares formula to find the best estimate of weights for each variable in the model, then loops through that a certain number of times until it converges to find the optimal weights, with the maximum number of iterations being defined as 100.

To assess independence, the threshold where a certain level of people dropped out was identified. In this case, a threshold is defined to be the minimum predictive probability where the algorithm labels a student as a graduate. After

https://doi.org/10.59720/25-166

quantifying the dropout rate in the training data, which was found to be around 39% (approximately 40%),the quantile in the male and female testing datasets was determined, where this proportion of students' predictive probabilities fell below this value, which outputted the threshold at which independence was satisfied. In other words, the independence threshold was calculated by finding the quantile corresponding to the observed dropout proportion in the overall training set.

To assess separation, the ROC curve was analyzed (Figure 2). In this case, the intersection between the male and female ROC curves was found, as well as the threshold at which this intersection occurred. This was done by interpolating the true positive and false positive rates and the thresholds and using the functions being output to identify where the true and false positive rates would be equal. The corresponding threshold was then calculated by determining the root of the difference between the interpolated functions, which gave the intersection value of the two ROC curves.

To assess sufficiency, the thresholds where the positive predictive value for males and females were equal were found. This value was determined through the positive predictive value of the model run on the dataset containing both the male and female students. The resulting value was around 71.6% and, from this, the thresholds where this was satisfied were calculated. This was done by interpolating the positive predictive values for each group as functions of their thresholds, then using these functions to find the threshold at which both males and females had the same positive predictive value. The overall results provided the true and false positive rates at the given thresholds to plot these points on the ROC curve. The threshold at which all three fairness criteria are satisfied is the threshold that must be used to create this hypothetical fair machine learning model.

Finally, the results from independence, separation, and sufficiency were combined in order to identify the threshold at which all three fairness criteria were satisfied.

REFERENCES

1. Angwin, Julia, *et al*. "Machine Bias." *ProPublica*. 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
2. Barocas, Solon, *et al*. "Fairness and Machine Learning: Limitations and Opportunities." *MIT Press*. 13 Dec 2023. https://fairmlbook.org/pdf/fairmlbook.pdf.
3. Gao, Jianhui, *et al*. "What Is Fair? Defining Fairness in Machine Learning for Health." *Wiley Online Library*, 15 Sept 2025. https://onlinelibrary.wiley.com/doi/10.1002/sim.70234.
4. Koçer, Rüya Gökhan, and Justyna Stypińska. "The Dynamic and Arbitrary Nature of Age as a Sensitive Characteristic in AI-Based Automated Decision Making Systems." *Springer Nature Link*, 22 May 2026. https://link.springer.com/article/10.1007/s43681-026-01096-1.
5. Kleinberg, Jon. "Inherent Trade-Offs in Algorithmic Fairness." *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 1, 17 Jan. 2019, p. 40. https://doi.org/10.1145/3308809.3308832.
6. Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review*, vol. 104, no. 3, 2016. https://www.cs.yale.edu/homes/jf/BarocasSelbst.pdf.
7. Flores, Anthony, *et al*. "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.'" *The Center for Justice and The Brennan Center for Justice*. http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf.
8. TheDevastator. "Predict Students' Dropout and Academic Success." Kaggle. https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention.
9. Gao, Jianhui, *et al*. "A Tutorial on Fairness in Machine Learning in Healthcare." *arXiv*, vol. 2406, no. 09307, 2024. https://arxiv.org/html/2406.09307v1#Sx4.T3.
10. Geneviève, Lester Darryl, *et al*. "Structural Racism in Precision Medicine: Leaving No One Behind." *BMC Medical Ethics*, 19 Feb. 2020. https://link.springer.com/article/10.1186/s12910-020-0457-8.
11. Office of the Privacy Commissioner of Canada. "Privacy Tech-Know Blog: When Worlds Collide – the Possibilities and Limits of Algorithmic Fairness (Part 1)." 5 April 2023 https://www.priv.gc.ca/en/blog/20230405_01/.

APPENDIX

GitHub Access Link: https://github.com/githubkshav/StudentPerformance

Received: January 26, 2025
Accepted: November 13, 2025
Published: September 04, 2026