PROJECT
(Professional Journal of English Education)
Volume 8, No. 5, September 2025
p–ISSN 2614-6320
e–ISSN 2614-6258

Analysis of Gen Z Preferences in Improving English Proficiency using Machine Learning

Keisha Naiza Djalle[1], Anita Anggraeni[2]

[1] Telkom University, Indonesia
[2] IKIP Siliwangi, Indonesia
[1] keishanaiza@student.telkomuniversity.ac.id, [2] anitaenglish26@ikipsiliwangi.ac.id

Abstract

This study analyzes Generation Z (Gen Z) preferences in enhancing English proficiency using machine learning with the Random Forest algorithm. The analysis aims to determine how accurately Random Forest can predict or classify models. The dataset contains three classes: "Applications," "Entertainment_Media," and "Tutoring." Data collected was processed during the pre-processing stage. With this dataset, the model was trained and identified the most influential features on model performance: factors, media, learning duration, understanding, and motivation. The model achieved an accuracy of 62% with hyperparameter tuning. This research aims to contribute to the development of more personalized learning methods for Gen Z.

Keywords: Gen Z; English Proficiency; Machine Learning

INTRODUCTION

Language is the primary communication tool used by humans in everyday life to convey information, ideas, and emotions. It is an innate ability possessed by humans as social beings. Language also refers to sound expressions produced by the movement of organs that are captured by the ear or auditory senses (Sari & Lestari, 2019). Without language, humans would be unable to interact with others. Especially in today's modern and digital era, strong language skills are highly valued in various fields of life. One important language used for communication is English. English is a critical language in the era of globalization. In addition to being an international language, English is a key requirement in education, employment, and access to digital information. In Indonesia, improving English proficiency is considered one of the key indicators for youth preparedness in facing global competition (Siregar & Wahyuni, 2022) . Generation Z those born between 1997 and 2012 is the most adaptive age group, as they were born amidst the digital revolution. This has enabled them to connect with various communication platforms that facilitate global communication (Sari, Ningsih, Pitri, Susmita, & Fazira, 2024). Therefore, it is essential for Gen Z to understand English. Although Gen Z is highly adaptive and grew up in a digital era, they have different preferences for enhancing their English skills. These variations suggest that a successful learning strategy for Generation Z is essential and should take into account their preferred methods of learning. To better comprehend these patterns, a machine learning-based strategy can be applied. Machine learning has already been widely applied in many fields, including education (Roihan, Sunarya, & Rafika, 2020). Many new educational opportunities have been made possible by technological advancements, especially in the area of machine learning (Sucianingtyas et al., 2025). Personalized learning, effective analysis, and more objective evaluation are just a few advantages of using machine learning in education (Hermawan et al., 2024). Furthermore, metrics like accuracy, precision, recall, and F1-score can be used to assess performance when machine learning is applied (Owoc et al., 2019). Machine learning also facilitates more adaptive

and effective language learning, as it can tailor learning activities to individual needs. By using this algorithm, Gen Z's learning preferences can be analyzed objectively, which can assist educational system designers in creating suitable learning strategies. This study aims to analyze Gen Z's preferences in improving English proficiency using machine learning and identify the factors influencing their choice of learning methods. Machine learning will be used to analyze the collected data in order to forecast Gen Z's preferences for better English. The Random Forest algorithm will be used to process the data. It is anticipated that this analysis will aid in the advancement of English language learning strategies and help educational establishments and online learning environments create more individualized curricula.

METHOD

A. Feature Design and Dataset

This study aims to analyze Gen Z's preferences in improving English proficiency using a machine learning approach. Data collection was conducted via an online questionnaire using Google Forms. The collected data consists of several features categorized as independent variables and one question serving as the label, reflecting whether participants like to write down new vocabulary or not. The dataset includes 445 respondents, with 150 using learning apps, 150 using entertainment media, and 145 attending online/offline tutoring.
The feature details for the dataset are outlined in Table 1:

Table 1. Dataset Features

Feature	Options
• Age **(age)**	• 13-17 • 18-22 • 23-28
• Gender **(gender)**	• Female • Male
• Current Education **(education)**	• Junior High/High School • Diploma (D1-D3) • Bachelor's Degree (D4/S1)
• Factor influencing the decision to improve English proficiency **(factor)**	• Engaging and not boring • Easy access • Affordable
• Motivation for learning English **(motivation)**	• Academic • Personal development and interest • Social and communication
• Frequency of learning English per week **(learnweek)**	• 1-2 • 3-4 • More than 4

• Time of day for learning English **(learntime)**	• Morning • Afternoon • Evening
• Duration of learning English **(learnlongtime)**	• 30 min-1 hour • 1-2 hours • More than 2 hours
• Motivation level for learning English **(howmotivated)**	• Scale of 1-5
• Obstacles in learning English **(obstacle)**	• Internet access • Lack of confidence • Lack of study time
• Understanding of English **(understanding)**	• Scale of 1-5
• Interest in writing down new vocabulary **(vocab)**	• Yes • No
• Main difficulty in learning English **(difficulty)**	• Listening • Speaking • Grammar/writing • reading
• Media used for learning English **(media)**	• Handphone • Laptop/computer • Notebooks

B. Random Forest

Random Forest is one of the popular machine learning algorithms due to its ability to handle complex data, minimize overfitting, and provide accurate predictions by combining multiple decision trees (Pratondo et al., 2023). Random Forest uses an ensemble of decision trees, built through bootstrap aggregating and random feature selection. The final prediction is made based on a majority voting mechanism across all trees. Another advantage of this algorithm is its ability to handle large and diverse datasets without the need for feature normalization or removal of extreme values (Pratondo & Bramantoro, 2022). Random Forest is also known for its ability to measure feature importance, which identifies the features most influential in the model's decisions.

In the training process, Random Forest not only takes random subsets of data but also selects random feature subsets for each node. This reduces overfitting significantly and improves generalization (Pratondo & Novianty, 2022). This happens because each tree is built by splitting the data, which is then trained with bootstrap samples involving 2/3 of the training data, while the remaining 1/3 is referred to as out-of-bag (OOB) and is used to calculate each tree's error rate (Indra et al., 2024). This characteristic of Random Forest makes it suitable for classifying data with many dimensions and non-linear feature correlations, such as the individual characteristics of Gen Z in selecting media for improving their English proficiency.

Model performance in Random Forest algorithms is largely determined by parameters. The number of decision trees is determined by the n_estimators parameter; in general, accuracy improves as the number of trees increases (Probst et al., 2019). The max_depth parameter limits

the maximum depth to prevent overfitting. Other parameters such as min_samples_split and min_samples_leaf set the minimum number of data points required to split a node, which can influence tree complexity (Liu et al., 2017). Next, there is max_features, which determines how many features are used at each node (Probst & Boulesteix, 2018). Moreover, there is bootstrap, which refers to the technique of sampling with replacement, helping to increase tree variation for better generalization (Abubakar et al., 2023). Methods like random search and grid search are also effective for optimizing parameter combinations (Rizky et al., 2024). Tuning these parameters has proven significant in handling class imbalance, thus understanding and optimizing them is key to the success of this algorithm.

C. Performance Metric

To evaluate the model, precise and pertinent performance metrics are needed. Accuracy, precision, recall, and F1-score are examples of performance metrics. There is also the confusion matrix, a table that displays the accuracy or inaccuracy of the model's predictions (Pratondo et al., 2023). True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are the four components that make up the confusion matrix. The performance of the model can be assessed using these elements. The percentage of accurate predictions compared to all predictions is known as accuracy. Accuracy is a useful metric for balanced datasets, but it is less useful for imbalanced datasets because it may exhibit high performance even when the model only recognizes the majority class (Grandini et al., 2020). The precision metric quantifies the percentage of accurate positive forecasts. This is crucial since false positives can have serious repercussions, like in the detection of diseases or fraud (Ferrer, 2022).. Meanwhile, recall measures the proportion of actual positives detected. Recall is crucial because missing positive cases presents a greater risk than misclassifying negatives (Saito & Rehmsmeier, 2015). F1-Score is a combined metric that balances precision and recall. This is especially useful when there is class imbalance, as it provides a comprehensive view of the model's performance on the minority class (Chicco & Jurman, 2020). All metrics come from the fundamental structure of the confusion matrix, and the choice of metrics depends on the problem context and needs (Tharwat, 2021).

RESULTS AND DISCUSSION

Results

After training and testing the model using the Random Forest algorithm with hyperparameter tuning, the results for the confusion matrix can be seen in Figure 1. This confusion matrix provides important information regarding model performance.

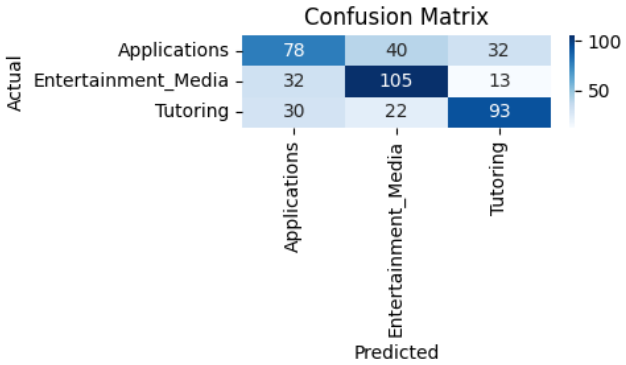

Figure 1. The confusion matrix

Random Forest is also known for its ability to show feature importance, or the features that contribute most to the model's performance. Feature importance assigns values to each feature and indicates how much each contributes to the model's decision-making process.
Feature importance is calculated based on impurity reduction. Gini impurity is one of the measures used to calculate uncertainty in the data. By understanding which features have high feature importance, we can identify the most influential ones. In this study, the top 5 features are shown in Table 2.

Table 2. Features

Feature	Importance
Factor	0.199
Media	0.092
Learnlongtime	0.077
Understanding	0.076
Motivation	0.065

Discussion

Based on the results, Figure 1 shows three labels: "Applications," "Entertainment Media," and "Tutoring." The results show that the True Positive (TP) predictions for "Applications" are 78, for "Entertainment_Media" are 105, and for "Tutoring" are 93. The overall model accuracy is 62%. The model is still not accurate enough to predict the labels.
For the "Applications" label, the precision is 55.7%, indicating that 55.7% of all predictions labeled as "Applications" are correct. Meanwhile, recall is 52%, showing that the model correctly identifies 52% of all actual "Applications" data. This indicates that the model often misclassifies "Applications" into other classes.
For the "Entertainment_Media" label, the model performs relatively better, with precision of 67.4% and recall of 64.1%. The model is fairly consistent in predicting and recognizing "Entertainment_Media" data. However, there are still some misclassifications.
For the "Tutoring" label, precision reaches 62.9% and recall is 70%. This indicates that the model is quite good at recognizing "Tutoring" data.
Overall, the model achieves a total accuracy of 62%. This shows there is room for improvement, particularly in distinguishing between "Applications" and "Entertainment_Media."
Based on the feature importance table in Table 2, the following is an explanation of the features that contribute most to the model's performance:

a. Factor (0.199): The factor influences how an individual, particularly Gen Z, views the learning process. The decision to learn is affected by how engaging and relevant the learning is.
b. Media (0.092): The media used also influences how the material or information is absorbed. The media must be tailored to individual habits and preferences.
c. Learnlongtime (0.077): The duration of learning reflects an individual's consistency in learning. The time spent learning correlates with how committed they are to improving their skills.
d. Understanding (0.076): The individual's understanding of English reflects their progress in learning. The better their understanding, the more capable they are of selecting and absorbing more complex material to improve their English proficiency.
e. Motivation (0.065): The motivation to learn English affects the level of dedication and focus during the learning process. A clear reason for learning helps an individual stay motivated to overcome challenges.

These five features have the most significant impact on the model's performance. By understanding these features, insights into Gen Z's preferences can be gained, helping to design

solutions that align with their learning preferences. Understanding these features also facilitates the development of more effective learning methods, utilizing appropriate technologies and approaches. Focusing on the most influential features can guide future learning strategies, improving engagement and success in learning English.

CONCLUSION

In this analysis, machine learning using the Random Forest algorithm was employed to predict Gen Z preferences in improving English proficiency. With labels "Applications," "Entertainment_Media," and "Tutoring," the model achieved an accuracy of 62%. This indicates that the model is still not sufficiently accurate in predicting the labels. Furthermore, the model found that the factors, media, learning duration, understanding, and motivation are the features that have the most significant influence on model performance. This analysis is expected to improve the development of more personalized English learning methods.

ACKNOWLEDGMENTS

We would like to thank everyone who helped me prepare this article.

REFERENCES

Abubakar, M. A., Muliadi, M., Farmadi, A., Herteno, R., & Ramadhani, R. (2023). Random Forest Dengan Random Search Terhadap Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung. *Jurnal Informatika*, *10*(1), 13-18.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1-13.

Ferrer, L. (2022). Analysis and comparison of classification metrics. *arXiv preprint arXiv:2209.05355*.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Hermawan, B. M., Hakim, M. A., Arifin, R., & Puspitasari, N. (2024, December). Pemanfaatan Artificial Intellegence, Khususnya Mechine Learning dan Deep Learning System dalam Pendidikan. In *Prosiding Seminar Nasional Amikom Surakarta* (Vol. 2, pp. 345-354).

Indra, D., Hayati, L. N., Daris, M. A., As' ad, I., & Mansyur, U. (2024). Penerapan Metode Random Forest dalam Klasifikasi Huruf BISINDO dengan Menggunakan Ekstraksi Fitur Warna dan Bentuk. *Komputika: Jurnal Sistem Komputer*, *13*(1), 29-40.

Liu, C. B., Chamberlain, B. P., Little, D. A., & Cardoso, Â. (2017). Generalising random forest parameter optimisation to include stability and cost. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10* (pp. 102-113). Springer International Publishing.

Owoc, M. L., Sawicka, A., & Weichbroth, P. (2019, August). Artificial intelligence technologies in education: benefits, challenges and strategies of implementation. In *IFIP international workshop on artificial intelligence for knowledge management* (pp. 37-58). Cham: Springer International Publishing.

Pratondo, A., & Bramantoro, A. (2022). Classification of Zophobas morio and Tenebrio molitor using transfer learning. *PeerJ Computer Science*, *8*, e884.

Pratondo, A., & Novianty, A. (2022, December). Comparison of wood classification using machine learning. In *2022 IEEE 10th Conference on Systems, Process & Control (ICSPC)* (pp. 308-312). IEEE.

Pratondo, A., Kurniawan, A. P., Eriyadi, M., Prasetyanto, F., Rahayu, D. P., & Akbar, M. D. (2023, August). Prediction of payment modes for online taxi users using machine learning. In *2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)* (pp. 411-415). IEEE.

Probst, P., & Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*(181), 1-18.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Rizky, M. H., Faisal, M. R., Budiman, I., Kartini, D., & Abadi, F. (2024). Effect of Hyperparameter Tuning Using Random Search on Tree-Based Classification Algorithm for Software Defect Prediction. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *18*(1), 95-106.

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan machine learning dalam berbagai bidang. *Jurnal Khatulistiwa Informatika*, *5*(1), 490845.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Sari, L., & Lestari, Z. (2019, February). Meningkatkan kemampuan berbicara bahasa Inggris siswa dalam menghadapi era revolusi 4.0. In Prosiding Seminar Nasional Program Pascasarjana Universitas PGRI Palembang.

Sari, M. N., Ningsih, P. E. A., Pitri, N., Susmita, N., & Fazira, S. (2024). Pentingnya Penguasaan Bahasa Bagi Gen Z. Jurnal Abdimas Adpi Sosial dan Humaniora, 5(3), 1-7.

Siregar, D. R., & Wahyuni, N. D. (2022). Urgensi penguasaan bahasa Inggris dalam menghadapi dunia kerja global. Jurnal Pendidikan dan Konseling, 4(3), 155–162.

Sucianingtyas, R., Falistya, L. R., Pujiana, S., Prayogi, A., & Laksana, S. D. (2025). Telaah Ragam Artificial Inteligence (AI) Dalam Pendidikan. *Madani: Jurnal Ilmiah Multidisiplin*, *3*(2), 232-243.

Tharwat, A. (2021). Classification assessment methods. *Applied computing and informatics*, *17*(1), 168-192.