World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Check for updates

Performance evaluation of random forest algorithm for automating classification of mathematics question items

Jubeile Mark Baladjay [1,*], Nisce Riva [1], Ladine Ashley Santos [1], Dan Michael Cortez [2], Criselle Centeno [1] and Ariel Antwaun Rolando Sison [1]

[1] Information Technology Department, Pamantasan ng Lungsod ng Maynila, Manila, Philippines.
[2] Computer Science Department, Pamantasan ng Lungsod ng Maynila, Manila, Philippines.

World Journal of Advanced Research and Reviews, 2023, 18(02), 034–043

Publication history: Received on 05 March 2023; revised on 27 April 2023; accepted on 29 April 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.18.2.0762

Abstract

Automated classification of mathematics question items based on the Table of Specifications is crucial in developing well-defined assessment content, significantly reducing teachers' workload. This study presents a performance evaluation of a Random Forest model designed to classify mathematics question items based on the content standards of the first quarter of tenth grade stipulated by the Philippines' Department of Education Curriculum Guide. The model uses an algorithm that extracts mathematical expressions as tokens for the Bag-of-words Model. The evaluation was conducted using precision, recall, F-1 score, and overall accuracy metrics, and the confusion matrix was used to assess the Random Forest model's performance. The results showed that the Random Forest model achieved 95% in precision, 95% in recall, 95% in F-1 score, and 95% in overall accuracy, demonstrating its effectiveness in classifying mathematics question items.

Keywords: Random Forest Algorithm; Bag-of-words; Confusion Matrix; Mathematical Information Retrieval; Table of Specification.

1. Introduction

Creating a meaningful student testing process through educational assessment requires using a Table of Specifications. (1) However, the classification of questions based on their topics, necessary for standardized assessment development, adds more work to the teaching personnel. Additional workloads and demands cause stress and negatively impact teaching performance. (2) Automating the tagging of question items can alleviate their burden. Although topic classifiers for sentences or phrases are abundant, the representation of mathematical expressions has posed a challenge in traditional Information Retrieval (3). Most natural language processing models omit mathematical symbols and numbers when extracting semantics from a sentence or corpus. Since mathematical expressions are prevalent in math questions, removing them from the corpus poses a challenge in classifying the topics of mathematics question items.

To address this issue, the proponents developed a Random Forest Model that can detect the topics of mathematics question items. This study focuses on classifying the topics of the Tenth Grade First Quarter Content Standards of the Philippines' Department of Education Curriculum Guide, specifically Series and Sequences, Polynomials, and Polynomial Equations. (4) This paper introduces a Bag-of-words Model that utilizes the Syntax Layout tree and transforms numbers and variables into generalized tokens to include mathematical expressions in a generalized manner. The Bag-of-words Model with tokenized mathematical expressions will be the features of the Random Forest Algorithm, which creates a model that classifies the topic of Mathematics Question Items. The Confusion Matrix will evaluate the model's performance in classifying each question item's topic.

* Corresponding author: Jubeile Mark A. Baladjay

2. Literature Review

In this section, the literature review of methods and techniques used in the methodology of this study.

2.1. Random Forest Algorithm

The Random Forest algorithm is a well-known ensemble machine learning method comprising multiple decision trees; The algorithm randomly chooses a subset of the node's candidate attributes. Then, it calculates the optimal splitting attribute for the node and uses bootstrapping aggregation, also known as the bagging method, to generate a set of decision trees. By randomly splitting the attributes of nodes, the algorithm creates diversity among each decision tree, which helps prevent identical outputs(5). Random Forest is a commonly used machine learning model for text classification, owing to its simplicity and superior performance compared to other models.(6) Researchers employed the XGBoost and Random Forest algorithms in a separate study for sentiment prediction on Twitter data. Their findings revealed that XGBoost outperformed Random Forest in terms of accuracy. The researchers then sought to improve the algorithms by incorporating additional parameters, enhancing accuracy scores(7).

2.2. Mathematical Language Processing

The role of mathematics semantics in language processing techniques is intriguing, as mathematics requires disambiguation for efficient data retrieval. By enabling the model to recognize the equivalence of mathematical formulas, the model can effectively distance equivalent formulas as zero, leading to more accurate and precise results in data retrieval. (8) Greiner-Petter et al. (9) suggest that utilizing Syntax Layout Tree to represent mathematical expression will be helpful for tokenization and semantic extraction.

Another way of extracting semantics is the identification of Identifiers and Definiens in mathematical formulae, which is crucial in facilitating information retrieval and comprehension. Identifiers are distinguished from other symbols by their unique meaning and significance within the context of a given formula. Meanwhile, Definiens refer to the textual descriptions of the identifiers based on the words surrounding them in the context. (10)

For example, Schubotz et al. (10) presented a case to illustrate this concept. In the mass-energy equivalence, "The relation between energy and mass is described by the mass-energy equivalence formula \mathbit{E}=\mathbit{m}\mathbit{c}^\mathbf{2}, where E is energy, m is mass, and c is the speed of light", E, m, and c are Identifiers. E represents energy, m represents mass, and c represents the speed of light, which are their respective Definiens pairs. This type of identification and definition enables more efficient information retrieval of complex mathematical concepts.

Meadows and Freitas (11) listed the inferential Spectrum of each Machine Learning Task from Extractive, closer to textual language processing, to more Abstract or symbolic in representation. They also concluded that these tasks still have constraints on varying levels of complexity.

2.3. Bag-of-Words Model

In Natural Language Processing or NLP field, the bag of words model is a common way of representing text data and a basic feature-based approach for text classification. It aims to label textual units automatically. It may target sentiment analysis, topic analysis, question answering, and natural language inference tasks. It involves representing text as numerical values. It can be used in topic modeling or topic classification. It produces embeddings that ignore the order, context, and semantic relations between words in a corpus. It encompasses statistics extracted from the corpus, such as the word frequency or the term Frequency-Inverse Document Frequency (TF-IDF) (12). Bag-of-words is one the most common text representation method for natural language processing. It has been notable in object categorization. (13)

Bag-of-words (BOW) is a simple text vectorizer technique for text mining. BOW creates a matrix of unique terms that can be expressed in one word or n-grams for all data instances. Term Frequency can be used to value each attribute in instances and terms. The formula to be used is as follows:

$$a_t(x) = TF_t(x) \tag{1}$$

Legends:

$a_t(x) = Attribute\ Value$
$t = term$

$x = document$
$TF = Term\ Frequency$
$TF_t(x) = the\ number\ of\ occurrences\ of\ term\ t\ in\ document\ x$

The matrix created by each attribute per document is the *document-term matrix* (14).

2.4. Confusion Matrix

In machine learning, the confusion matrix is an absolute term (15). It is present in different areas such as computer vision, natural language processing (NLP), acoustics, and more, especially in evaluating scientific models and engineering applications. (16) The confusion matrix assesses the model's accuracy by comparing predicted and actual values. (15) A cross table keeps track of the number of occurrences between the two raters, the actual classification, and the predicted classification. It represents the percentages of four possible classification outcomes such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN). (17)

Suppose the number of correct positives classified as positive is called True Positive. Suppose the number of actual negatives correctly classified as unfavorable is called True Negative if the number of actual positives incorrectly classified as unfavorable is called False Negative. If the number of actual negatives that are incorrectly classified as positives is called False Positive. (15)

Precision is the percentage of units our model predicts will be Positive and Positive. It says how much we can rely on the model when it predicts a person as Positive. (17)

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall measuring the predictive accuracy for the positive class of the model. It measures the ability of the model to find all the positive units in the dataset.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Accuracy measures how much the model is correctly predicting in the entire dataset. It assumes values between 0 and 1. The quantity missing to reach one is called Misclassification Rate.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

F1-Score assesses the model's performance classification starting from the confusion matrix. In terms of Multi-class cases, it should involve all the classes.

$$F1 - Score = \left(\frac{2}{precision^{-1} + recall^{-1}}\right) = 2 \cdot \left(\frac{precision * recall}{precision + recall}\right) \tag{5}$$

3. Methodology

In this section, the proponents will discuss the research development process of the Random Forest Model. It has six (6) phases: Problem Definition, Data Collection, Data Pre-processing, Feature Extraction, Random Forest Algorithm Training and Testing of the Dataset, and Model Evaluation. The flow is displayed in Figure 1.

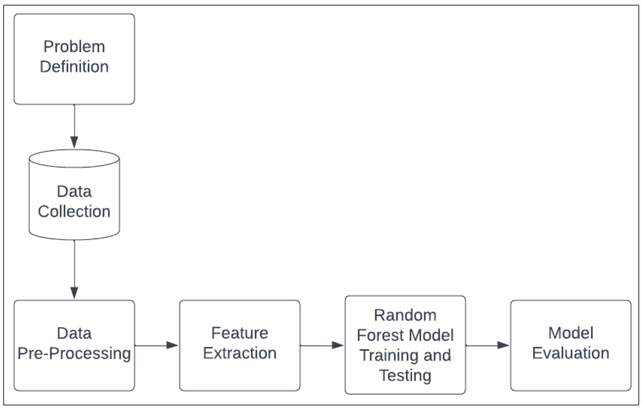

Figure 1 Research Development Process

3.1. Problem Definition

The proponents identified the scope of their classifying algorithm. The proponents used the Curriculum Guide to Identify the topics to be classified by assessing the curriculum to segregate topics into three labeled topics: Series and Sequences, Polynomials, and Polynomial Functions. The segregated topics were consulted with Math Professionals for verification if the proponents were valid labels for mathematics question items. The scope of the topics will be for Grade 10 1st Quarter only. The label formation was based on the content standards set by the Philippines Department of Education (DepEd); an adaptation of the table is shown in Table 1.

Table 1 Label formation of classification based on Content Standards

Content: Patterns and Algebra	
Content Standards: The learner demonstrates an understanding of key concepts of sequences, polynomials, and polynomial equations.	
Label	**Learning Competency**
Series and Sequences	Generates Pattern Illustrates Arithmetic Sequence Determines arithmetic means and nth term of an arithmetic sequence Finds the sum of the terms of a given arithmetic sequence. Illustrates a Geometric Sequence Differentiates a Geometric Sequence from an Arithmetic Sequence Differentiates a Finite Geometric Sequence from an infinite Geometric Sequence Determines geometric means and nth term of a geometric sequence. Finds the sum of the terms of a given finite or infinite geometric sequence. Illustrates other types of sequences (e.g., harmonic, Fibonacci). Solves problems involving sequences.
Polynomials	Performs division of polynomials using long division and synthetic division. Proves the Remainder Theorem Factor Theorem. Factors polynomials. Solves Problems involving polynomials
Polynomial Equation	Illustrates polynomial equations. Proves Rational Root Theorem. Solves Polynomial Equation Solves Problem Involving Polynomial Equations

Table Adapted from the K to 12 Curriculum Guide in Mathematics released by the Philippines Department of Education. (4)

3.2. Data Collection

The proponents gathered multiple datasets labeled or manually labeled by the proponents. The proponents extracted a dataset from DeepMind's published Mathematics Dataset composed of the.txt files in a Freeform Question/Answer Format. (18) The proponents gathered multiple questions for the Series and Sequences and Polynomial Equation Dataset. The proponents also manually typed into LaTeX expressions gathered from DepEd Grade 10 Modules. A total of 924 instances of Dataset were Acquired.

3.3. Data Pre-Processing

The proponents translated the LaTeX Expression dataset into Freeform Questions to be easily tokenized for extracting Features and Rendering Models.

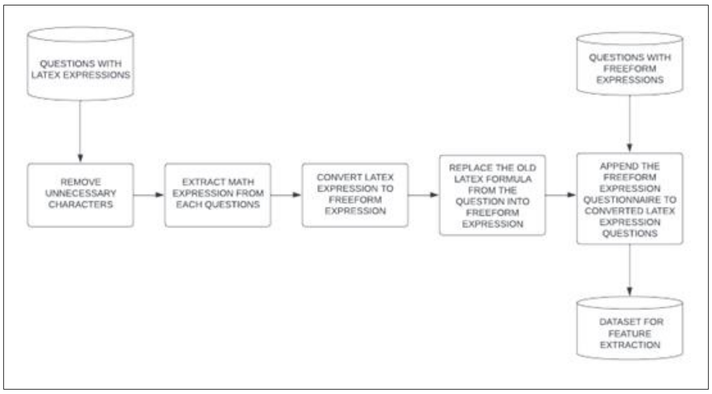

Figure 2 Translation of LaTeX Encoded Mathematical Expressions to freeform expression

In this paper, Mathematics expression ambiguity relies on multiple math expressions representing one topic at a time, whether for series and sequences, polynomials, or polynomial equations. In this study, equations are translated to their Expression Tree Syntax to convey a similar structure to other equations since they were already labeled. The similarity of equations and their other features may define their label.

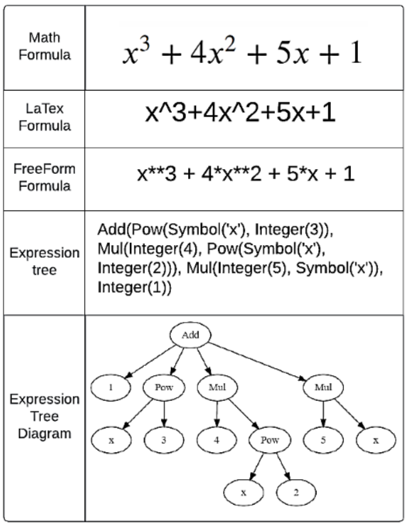

Figure 3 Visualization of Math Formula in LaTeX Format, FreeForm Format, Expression Tree Format, and Its Expression Tree Diagram

The researchers used SymPy and LaTeX to SymPy python libraries to represent the LaTeX Equations and Freeform Equations as Expression Trees (19), remove numbers, and generalize each symbol to x to represent a single meaning across other questions that uses other alphabetical characters in their questions.

3.4. Feature Extraction

For **Feature Extraction**, the proponents tokenize words and Each Expression as one token representing the feature. The proponents used the bag of words technique to produce features and are used to vectorize a question to be predicted.

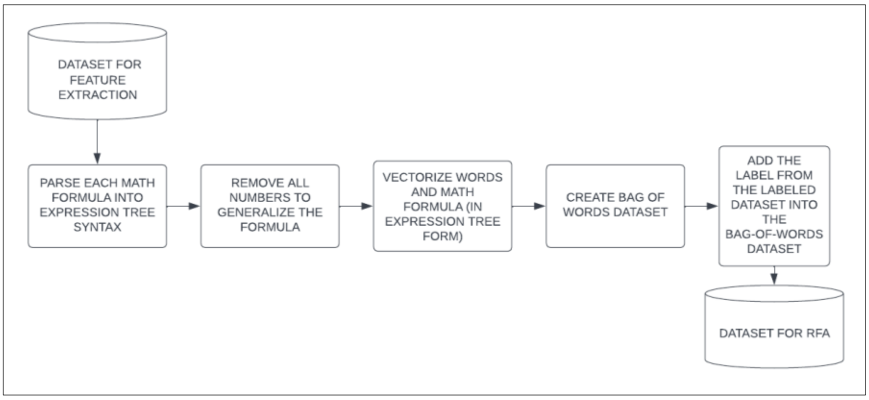

Figure 4 Generalization of Variables and Numbers in Expression Tree

The researchers will utilize bag-of-words as the feature extraction technique for creating the Model for the Random Forest Algorithm, where words and math expressions are treated as specific terms.

3.5. Random Forest Training and Testing

The proponents tune the model into a 75-25 split of test and train data, tune the number of estimators to 20, and set the randomness to 0.

Random forest is a high-performing supervised learning method that is flexible and easy to utilize, as it has also been effective in classification (20). It comprises multiple decision trees during training, produces output by evaluating individual trees, and has more efficiency than decision trees(21). Random Forest is an extension of the bagging technique as an ensemble algorithm. Each decision tree is built by randomly selecting subsets of features from the dataset. The training data is then obtained by applying the bootstrap method. In identifying an unrecognized value, each decision tree gives a single vote, in which the class with the most votes is decided (22).

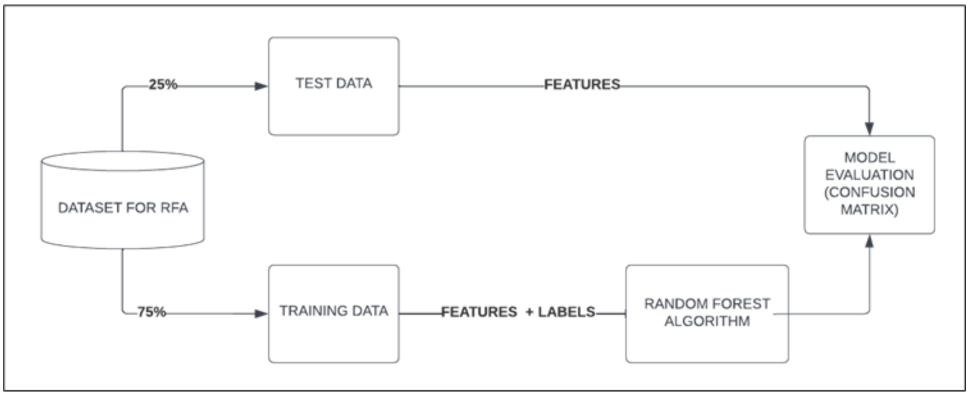

Figure 5 Random Forest Algorithm Training and Testing

3.6. Model Evaluation

The proponents utilized a confusion matrix to determine the generated performance of the model's overall accuracy, precision, recall, and f1-score.

4. Results and Discussion

The proponents formulated a dataset with features that will identify if it is the designated topic derived from the cleaned data. The new dataset with features was trained and tested using Random Forest Algorithm using Sci-Kit Learn packages. The model achieved 95% accuracy in predicting questions if it falls under series and sequences, polynomials, and polynomial equations with a 75-25 train test split.

Table 2 Confusion Matrix Table for Actual and Predicted Values

Class		**Prediction**			**Total (actual)**
		Polynomial equation	Polynomials	Series and sequences	
Actual	Polynomial equation	77	2	3	82
	Polynomials	6	79	0	85
	Series and sequences	2	0	70	72
TOTAL (prediction)		85	81	73	239

Table 2 displays the confusion matrix of 25% of the testing data. The model correctly classified 77 instances of Polynomial Equation Class, three (3) instances classified as Polynomials, and two (2) instances are Series and Sequences. The model predicted 86 instances of the Polynomial Equation class with only 82 actual instances. In Polynomial Class, the model correctly predicted 78 instances; seven (7) instances were classified as "Polynomial Equation," and No instance of the Polynomials class was classified as Series and Sequence. The model predicted 81 instances of Polynomials, with only 85 actual instances. Lastly, the model correctly classified 70 instances of the Series and Sequences class with two (2) instances classified as Polynomial Equation and No Series and Sequences misclassified as Polynomials. The Individual Label tally is presented in succeeding tables (Table 3, Table 4, Table 5)

Table 3 Tally of Actual Vs. Predicted Classification on Polynomial Equation Class

Polynomial Equation	**Predicted True**	**Predicted False**	**Total (Actual)**
Actual True	77	5	82
Actual False	9	148	157
Total (Prediction)	86	153	239

Table 4 Tally of Actual Vs. Predicted Classification on Polynomials Class

Polynomials	**Predicted True**	**Predicted False**	**Total (Actual)**
Actual True	79	6	85
Actual False	2	152	154
Total (Prediction)	81	158	239

Table 5 Tally of Actual Vs. Predicted Classification on Series and Sequences Class

Series and Sequences	**Predicted True**	**Predicted False**	**Total (Actual)**
Actual True	70	2	72
Actual False	2	165	167
Total (Prediction)	72	167	239

Table 6 Accuracy, Precision, Recall, F1-Score, and Support

Class	Accuracy	Precision	Recall	F1-score	Support
Polynomial equations	94%	91%	94%	92%	82
Polynomials	96%	98%	93%	95%	85
Series and sequence	98%	96%	97%	97%	72
Accuracy				95%	239
Macro average		95%	95%	95%	239
Weighted average		95%	95%	95%	239

Table 6 displays the testing set results that the model achieved a 94% accuracy rate for classifying Polynomial Equations, with 91% precision, 94% recall, and a 92% f1-score based on 82 instances from the dataset. The Polynomials class results also showed a high accuracy rate of 96%, with 98% precision, 93% recall, and a 95% f1-score based on 85 instances. The model achieved a 98% accuracy rate for Series and Sequences, with 96% precision, 97% recall, and a 97% f1-score based on 72 instances. Overall, the model's accuracy rate of 95% and 95% macro and the weighted average for all other criteria, including precision, recall, and f1-score, demonstrate its outstanding performance in classifying three topics based on the content standard of the tenth-grade first quarter of the Philippines Mathematics Curriculum Guide, using the Random Forest Algorithm.

5. Conclusions

The model's performance in classifying mathematics question items has exceeded the outstanding benchmark, achieving a 95% accuracy rate and a 95% macro and weighted average precision, recall, and f-1 score, indicating its high predictive power. The model achieved high scores of over 90% in all criteria for individual topic classification, which could be valuable for teachers' question banks in classifying topics for tenth-grade students. The study is a foundation for more detailed classifications based on the Table of Specifications' Bloom's Taxonomy label and Learning Competency topics for questions. However, the study is limited to the tokens in the dataset, and words not included in the dataset may need to be misclassified as one of the three topics. The proponents suggest using a more abundant and diverse data set to improve the study.

Compliance with ethical standards

Acknowledgments

The Authors would like to thank Pamantasan ng Lungsod ng Maynila College of Engineering for supporting this academic endeavor.

Disclosure of Conflict of interest

The Authors proclaim no conflict of interest.

References

[1] Osebhohiemen E. Use of Table of Specification in Construction of Teacher-Made Achievement Test in Mathematics in the Primary and Secondary Schools. Vol. 5, Melting Pot. 2019.

[2] Sarabia A, Collantes LM. Work-Related Stress and Teaching Performance of Teachers in Selected School in the Philippines. IRJE |Indonesian Research Journal in Education [Internet]. 2020;4(1):6–27. Available from: https://online-journal.unja.ac.id/index.php/irje/index

[3] Zhong W, Lin J. PYA0: A Python Toolkit for Accessible Math-Aware Search. In: SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, Inc; 2021. p. 2541–5.

[4] K to 12 Curriculum Guide MATHEMATICS [Internet]. 2016. Available from: http://lrmds.deped.gov.ph/.

[5] Li X, Wang Z, Wang L, Hu R, Zhu Q. A multi-dimensional context-aware recommendation approach based on improved random forest algorithm. IEEE Access. 2018 Aug 13;6:45071–85.

[6] Jalal N, Mehmood A, Choi GS, Ashraf I. A novel improved random forest for text classification using feature ranking and optimal number of trees. Journal of King Saud University - Computer and Information Sciences. 2022 Jun 1;34(6):2733–42.

[7] Supriya BN, Akki CB. Sentiment prediction using enhanced xgboost and tailored random forest. International Journal of Computing and Digital Systems. 2021;10(1):191–9.

[8] Schubotz M, Greiner-Petter A, Scharpf P, Meuschke N, Cohl HS, Gipp B. Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. Institute of Electrical and Electronics Engineers Inc.; 2018. p. 233–42.

[9] Greiner-Petter A, Youssef A, Ruas T, Miller BR, Schubotz M, Aizawa A, et al. Math-word embedding in math search and semantic extraction. Scientometrics. 2020 Dec 1;125(3):3017–46.

[10] Schubotz M, Grigorev A, Leich M, Cohl HS, Meuschke N, Gipp B, et al. Semantification of identifiers in mathematics for better math information retrieval. In: SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, Inc; 2016. p. 135–44.

[11] Meadows J, Freitas A. A Survey in Mathematical Language Processing. 2022 May 30; Available from: http://arxiv.org/abs/2205.15231

[12] Souza FD, Filho JB de O e. S. Embedding generation for text classification of Brazilian Portuguese user reviews: from bag-of-words to transformers. Neural Comput Appl. 2022;

[13] Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 2010 Dec;1(1–4):43–52.

[14] Cichosz P. A case study in text mining of discussion forum posts: Classification with bag of words and global vectors. International Journal of Applied Mathematics and Computer Science. 2018;28(4):787–801.

[15] Zeng G. On the confusion matrix in credit scoring and its analytical properties. Commun Stat Theory Methods. 2020 May 2;49(9):2080–93.

[16] Krstinić D, Braović M, Šerić L, Božić-Štulić D. Multi-label Classifier Performance Evaluation with Confusion Matrix. In Academy and Industry Research Collaboration Center (AIRCC); 2020. p. 01–14.

[17] Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. 2020 Aug 13; Available from: http://arxiv.org/abs/2008.05756

[18] Saxton D, Grefenstette E, Hill F, Kohli P. ANALYSING MATHEMATICAL REASONING ABILITIES OF NEURAL MODELS. In: International Conference on Learning Representations [Internet]. 2019. Available from: https://github.com/deepmind/mathematics_dataset

[19] Meurer A, Smith CP, Paprocki M, Čertík O, Kirpichev SB, Rocklin M, et al. SymPy: Symbolic computing in python. PeerJ Comput Sci. 2017;2017(1).

[20] Huljanah M, Rustam Z, Utama S, Siswantining T. Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. In: IOP Conference Series: Materials Science and Engineering. Institute of Physics Publishing; 2019.

[21] Vigneswari D, Kumar N, Raj V, Gugan A, Vikash SR. Machine Learning Tree Classifiers in Predicting Diabetes Mellitus. 2019.

[22] Liu Y, Liu C, Shen Y, Zhao X, Gao S, Huang X. Non-intrusive energy estimation using random forest based multi-label classification and integer linear programming. Energy Reports. 2021 Nov 1;7:283–91.

Author's short biography

	**Mr. Jubeile Mark A. Baladjay:** He is a 4th Year Bachelor of Science in Information Technology student at Pamantasan ng Lungsod ng Maynila. He is a Web Developer Volunteer at Google Developers Student Clubs - PLM. His research interests are machine learning, artificial intelligence, data science, and web development technologies.
	**Mr. Nisce N. Riva:** He is a 4th Year Bachelor of Science in Information Technology student at Pamantasan ng Lungsod ng Maynila. His research interests are machine learning, artificial intelligence, data science, and web development technologies
	**Ms. Ladine Ashley L. Santos:** She is a 4th year Bachelor of Science in Information Technology student at Pamantasan ng Lungsod ng Maynila. She is also a Junior Level Science Scholarship Merit scholar of the Department of Science and Technology Philippines. Her interest mainly lies in web development and software engineering technologies.
	**Dr. Dan Michael A. Cortez:** He is currently an Associate Professor and the Program Chair of the Computer Science Department with ten (10) years of teaching experience. He graduated with a Bachelor of Science in Information Technology and also obtained his Master of Science in Information and Communications Technology degree from the Pamantasan ng Lungsod ng Maynila. He finished his Doctor in Information Technology from the Technological Institute of the Philippines-Quezon City Campus. He is also an author of various books and has already published his research in the field of Information Technology, both locally and internationally.
	**Prof. Criselle J. Centeno:** She is a full-time faculty of Pamantasan ng Lungsod ng Maynila. She has 14 years of teaching experience. She graduated from Quezon City University with a Bachelor of Science in Information Technology and a Master of Information Technology at the Technological University of the Philippines. She is pursuing her Doctor in Information Technology at the Laconsolacion University of the Philippines - Malolos Bulacan Campus.
	**Mr. Ariel Antwaun Rolando C. Sison:** He has five years of teaching experience. He graduated from Emilio Aguinaldo College with a Bachelor of Science in Information Technology. He is pursuing a Master in Information Technology at the Technology Institute of the Philippines - Quezon City (TIP-QC).