Multiple Choice Question Difficulty Level Classification with Multi Class Confusion Matrix in the Online Question Bank of Education Gallery

Pariang Sonang Siregar[1,*], Rindi Genesa Hatika [2], B. Herawan Hayadi [3]

[1] Department of Elementary Teacher Education, Universitas Rokania, Indonesia

[3] Department of Physics Education, Universitas Pasir Pengaraian, Indonesia

[3] Information Technology Education, Universitas Bina Bangsa, Indonesia

(Received: October 6, 2023; Revised: November 16, 2023; Accepted: November 25, 2023; Available online: December 5, 2023)

Abstract

This study aims to enhance the quality of education by addressing the crucial aspect of test question planning, specifically focusing on the identification of item difficulty as a classification problem. The research method involves the utilization of various machine learning methods, including Random Forest, Logistic Regression, SVM, Gaussian, and Dense Neural Network, with an exploration of embedding, lexical, and syntactic features. The dataset comprises questions from elementary and junior high school levels. The primary purpose is to provide teachers with a tool to predict and categorize the difficulty level of test questions as easy, medium, or difficult, thereby aligning assessments with students' abilities. Our research contributes to the field by framing item difficulty identification as a classification problem and systematically evaluating multiple machine learning methods. The findings reveal that Random Forest emerges as the most effective method, achieving an accuracy of 84% in subjects and 80% in other cases. This highlights the practical applicability of machine learning in assisting educators in crafting assessments that accurately reflect students' comprehension levels. Furthermore, the study demonstrates that the incorporation of feature embedding and TF-IDF significantly enhances the accuracy of the resulting model. This insight into effective feature utilization contributes valuable knowledge for future research in educational assessment and machine learning applications in education.

Keywords: Education Quality Enhancement, Machine Learning Methods, Item Difficulty Classification, Feature Embedding and TF-IDF

1. Introduction

Education is a key aspect in community development and human resource development. One of the commonly used learning evaluation methods is multiple-choice based exams [1]. In today's digital era, multiple-choice exams are often organized through Education Gallery's Online Question Bank platform [2]. In this context, it is important to ensure that the difficulty level of multiple-choice exam questions presented to students is appropriate for their ability level [3].

Determining the difficulty level of exam questions is a challenge in online education management. Therefore, this research aims to develop an automatic classification system that can assess the difficulty level of multiple-choice questions in the Education Gallery Online Question Bank. This system is expected to help teachers and instructors in designing exams that match students' abilities, improve the validity of learning evaluations, and provide students with a better learning experience.

The importance of education in developing quality human resources cannot be doubted. Through the education process, humans acquire essential knowledge to develop various aspects of life [4]. Good education involves learning evaluation at all levels, one of which is through examinations [5]. In this evaluation process, the creation of exam questions is very important to measure students' abilities and talents. The design of questions must consider the level of difficulty in order to distinguish the diverse ability levels of students. The identification of question difficulty is important, and is usually divided into three types: easy, medium, and difficult.

***Corresponding author: Pariang Sonang Siregar (pariangsonangsiregar@gmail.com)**

DOI: https://doi.org/10.47738/jads.v4i4.132

In this context, Natural Language Processing (NLP) can be a solution to the problem of identifying the difficulty level of questions using text classification techniques. Text classification is the process of grouping documents into categories or classes based on their content [6]. Text classification approaches help in recognizing the difficulty level of questions by predicting unknown variables based on other variables [7]. Previous research has shown that prediction of the difficulty level of high-stakes medical exam questions can be done well using embedding, followed by linguistic features using Random Forest algorithm [8,9]. Another study used Item Response Theory (IRT) to predict the difficulty level of yes/no questions. Although there were some prediction errors, this study tried using synonyms or changing the question to increase the difficulty level of the question [10].

In this study, various models such as Random Forest, SVM, Gaussian, and Dense Neural Network (DNN) were used to identify the difficulty level of questions in subjects [11]. In addition, the Logistic Regression method was also added for experimentation. This study predicts the difficulty level of questions by using three types of feature extraction: embedding features using Word2Vec, lexical features using TF-IDF, and syntactic features using POS tagging method. To generate difficulty level labels (easy, medium, difficult), a survey was conducted among elementary and junior high school teachers. The data was then processed to fit the needs during modeling.

This research proposes the use of classification method with Multi Class Confusion Matrix to classify multiple-choice questions into different difficulty levels. By utilizing technology and data analysis, this research aims to improve the precision and accuracy in assessing the difficulty level of exam questions. It is expected that the results of this research can make a positive contribution to the development of online education, help improve the quality of learning evaluation, and provide more relevant information to teachers and students in the teaching and learning process.

2. Literature Review

2.1. E-Education

E-Education, or electronic education, is a concept related to the use of information and communication technology (ICT) in an educational context [12]. The theory of E-Education includes the use of the internet, educational software, and computer hardware to enhance the learning experience. The main goal of E-Education is to bring flexible, interactive, and affordable learning to students around the world. One important aspect of E-Education theory is accessibility. Through the use of technology, students have access to various learning resources, learning modules, and instructors without having to be in the same physical location. This allows students from different geographical, economic, and social backgrounds to access education without time and space barriers.

In addition, E-Education also facilitates the use of innovative learning methods. Through E-Education platforms, teachers can integrate multimedia, simulation, and interactive elements into the learning experience [13][14][15]. This helps create an engaging learning environment and motivates students to engage in learning. E-Education theory also focuses on the development of digital skills. In this digital age, skills such as digital literacy, the ability to sort information, and the ability to communicate through digital media are essential. E-Education helps students develop these skills early on, preparing them for success in a technology-based society. In addition to the benefits for students, E-Education also strengthens the teacher's role as a learning facilitator. Teachers can monitor students' progress more efficiently, provide immediate feedback, and design learning tailored to students' individual needs. By understanding E-Education theory, educators can create dynamic learning experiences, combining pedagogical principles with modern technology to achieve optimal learning outcomes.

Understanding the level of difficulty in the E-Education era is crucial, as this approach allows for effective personalization of learning. In an E-Education environment, each student can have different levels of understanding and ability. By understanding the level of difficulty, educators can design learning content that suits the level of readiness and individual needs of students. This not only improves students' understanding but also maintains their motivation in the learning process. In addition, understanding the level of difficulty also helps in optimizing the use of technology. By customizing learning materials according to the level of difficulty, educators can choose the most suitable apps, platforms, or learning methods to help students understand concepts better. Thus, understanding the level of difficulty not only leads to improved learning outcomes but also ensures that students' learning experiences in E-Education environments become more efficient, effective, and meaningful. E-Education, or electronic education, is a concept related to the use of information and communication technology (ICT) in an educational context [12]. The theory of E-Education includes the use of the internet, educational software, and computer hardware to enhance the learning experience. The main goal of E-Education is to bring flexible, interactive, and affordable learning to students around the world. One important aspect of E-Education theory is accessibility. Through the use of technology, students have access to various learning resources, learning modules, and instructors without having to be in the same physical location. This allows students from different geographical, economic, and social backgrounds to access education without time and space barriers.

In addition, E-Education also facilitates the use of innovative learning methods. Through E-Education platforms, teachers can integrate multimedia, simulation, and interactive elements into the learning experience [13][14][15]. This helps create an engaging learning environment and motivates students to engage in learning. E-Education theory also focuses on the development of digital skills. In this digital age, skills such as digital literacy, the ability to sort information, and the ability to communicate through digital media are essential. E-Education helps students develop these skills early on, preparing them for success in a technology-based society. In addition to the benefits for students, E-Education also strengthens the teacher's role as a learning facilitator. Teachers can monitor students' progress more efficiently, provide immediate feedback, and design learning tailored to students' individual needs. By understanding E-Education theory, educators can create dynamic learning experiences, combining pedagogical principles with modern technology to achieve optimal learning outcomes.

Understanding the level of difficulty in the E-Education era is crucial, as this approach allows for effective personalization of learning. In an E-Education environment, each student can have different levels of understanding and ability. By understanding the level of difficulty, educators can design learning content that suits the level of readiness and individual needs of students. This not only improves students' understanding but also maintains their motivation in the learning process. In addition, understanding the level of difficulty also helps in optimizing the use of technology. By customizing learning materials according to the level of difficulty, educators can choose the most suitable apps, platforms, or learning methods to help students understand concepts better. Thus, understanding the level of difficulty not only leads to improved learning outcomes but also ensures that students' learning experiences in E-Education environments become more efficient, effective, and meaningful.

2.2. Difficulty Level Classification

Difficulty classification is an important theory and technique in the field of machine learning that aims to categorize objects or data into certain classes or categories based on their characteristics or attributes [11][16][17]18]. In the context of this research, difficulty classification focuses on grouping exam questions into three main categories: easy, medium, and difficult. The main purpose of the difficulty classification theory is to assist teachers and evaluators in understanding the difficulty level of each question and, therefore, help them design exams that are balanced and appropriate for students' abilities.

Difficulty classification methods use various machine learning algorithms, such as Random Forest, SVM, Naive Bayes, and Logistic Regression, to identify patterns and relationships in the data that describe the difficulty level of questions. For example, using the Random Forest algorithm, the model can take into account various attributes or features of the question, such as embedding, lexical features, and syntactic features, to make accurate predictions about the difficulty level of each question.

At the application level, difficulty classification has various important applications in the field of education. Teachers can use the classification results to adjust their teaching and evaluation methods according to students' ability levels. In addition, difficulty classification also provides valuable insights for curriculum developers, helping them to structure learning materials that match the expected difficulty level. Therefore, the theory of difficulty classification not only refers to machine learning algorithms and techniques but also includes its practical applications that support more effective and targeted educational decision-making. By understanding and implementing this theory, educators can optimize students' learning experiences and improve the overall quality of education.

2.3. Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between humans and computers through human language [19]. The main goal of NLP is to enable computers to understand, interpret, and respond to human language in a meaningful way. In its theoretical underpinnings, NLP covers the set of grammatical, syntactic, semantic, and pragmatic rules that govern the structure of language. NLP has two main approaches: rule-based approaches and statistical-based approaches. The rule-based approach uses human grammatical rules and structures to understand language, while the statistical-based approach uses machine learning techniques to predict words or structures based on statistics from training data.

One of the key concepts in NLP is context understanding. Human language is highly contextual; that is, the meaning of a word or phrase depends on the surrounding words or phrases. To address this complexity, modern NLP models use techniques such as Word Embeddings and Neural Networks to represent words in a vector space that considers the semantic relationships between them. Word Embeddings, such as Word2Vec and GloVe, convert words into numerical vectors so that computers can understand the semantic relationships between them.

The importance of context also leads to another field in NLP called natural language understanding (NLU). NLU involves understanding the context and meaning behind words and sentences. Techniques like Named Entity Recognition (NER) and Sentiment Analysis are examples of powerful NLP applications in NLU. NER helps identify named entities such as names of people, places, and organizations in text, while Sentiment Analysis is used to assess the sentiment or feelings contained in text, for example, whether a review is positive or negative.

In recent years, the development of Deep Learning has also given a huge boost to the advancement of NLP. Models such as Recurrent Neural Networks (RNN) and Transformers have been used to model complex contextual relationships in text. Transformers, in particular, introduced through models such as GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), have solved several difficult problems in NLP, including language understanding, translation, and text generation. As technology continues to evolve, NLP continues to play an important role in everyday applications, from virtual assistants to sentiment analysis and automatic translation.

2.4. Machine Learning

Machine Learning (ML) is a branch of artificial intelligence that allows computer systems to learn and make decisions from data without having to be explicitly programmed [2]-[6][20]. The basic principle behind machine learning is the ability of computers to recognize complex patterns in data and use this understanding to make intelligent predictions or decisions. There are several types of machine learning, but the most common are supervised learning, unsupervised learning, and reinforcement learning.

2.5 Supervised Learning

In supervised learning, machine learning models are trained using datasets that contain matched inputs and outputs [21]. The main goal is to teach the model to understand the relationship between inputs and outputs so that it can make accurate predictions when faced with new data that it has never seen before. Examples of supervised learning applications are in classification (categorizing data into predefined classes) and regression (predicting values based on input data).

2.6. Unsupervised Learning

Unsupervised learning involves clustering data without the use of pre-matched labels or outputs [22][23]. Unsupervised learning models are designed to find intrinsic patterns or structures in data. Algorithms in unsupervised learning try to group data into clusters or categories based on the similarity of certain features or characteristics.

2.7. Reinforcement Learning

Reinforcement learning is a machine learning paradigm in which an agent learns how to create a sequence of actions based on interactions with an environment to achieve a specific goal [24]. In order for the agent to achieve the goal, it is given positive (reward) or negative (penalty) feedback based on the decisions or actions it takes. The goal of reinforcement learning is to optimize the actions taken by the agent so that the total reward obtained by the agent from its environment is maximized.

2.8. Feature Extraction and Selection

In machine learning, feature selection and feature extraction are very important [25][26]. Feature selection is the process of selecting the most relevant and influential subset of features from the entire dataset, while feature extraction involves transforming raw data into a more meaningful form, which often reduces the dimensionality of the data but retains important information. Smart feature selection and proper feature extraction can improve model performance and reduce data dimensionality, which is often a challenge in complex data processing. With these techniques, machine learning enables systems to understand patterns and relationships in very large and complex data, providing the ability to make intelligent decisions and predict outcomes with high accuracy. Machine learning has found applications in areas such as facial recognition, text analysis, medical diagnostics, autonomous vehicles, and more, proving its importance in the modern data-driven world.

3. Methodology

3.1. Research Stages

This study aims to apply classification methods to predict the difficulty level of questions and analyze them. We use a dataset of questions from elementary and junior high school practice questions. Several methods will be tested, and their performance evaluated. The methods used in this question text classification include Random Forest, SVM, Naive Bayes, Logistic Regression, and DNN. Random Forest consists of many decision trees that are used for prediction. SVM uses the best hyperplane between classes. Naive Bayes uses Bayes' theorem. Logistic Regression predicts probabilities. DNN uses an artificial neural network with many hidden layers connected between each neuron. Here is the design of the system built to classify questions, as shown in Figure 1 below.

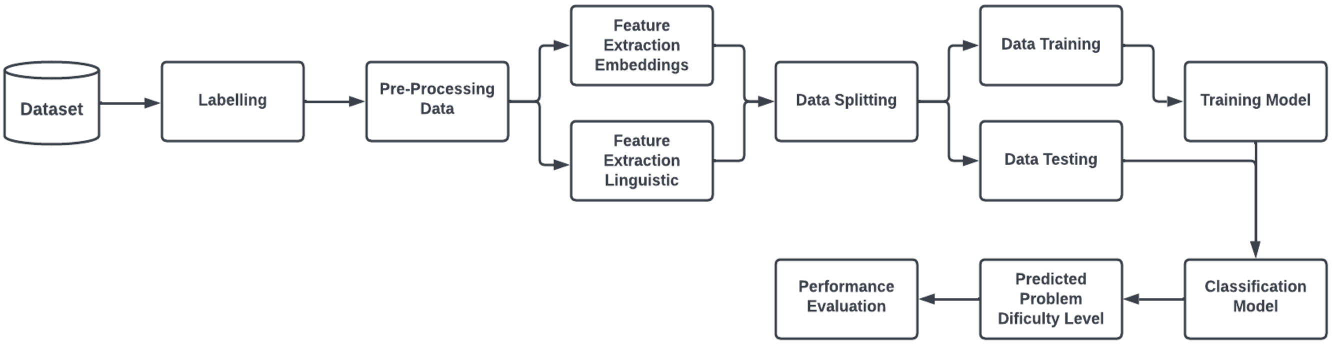

Figure 1. Research flow

Figure 1 delineates the detailed process of multiple-choice question classification. The journey commences with the assembly of a dataset comprising a total of 548 question entries. This dataset then undergoes a meticulous manual labeling process, facilitated through a survey administered to teachers, ensuring the assignment of appropriate labels. Following the labeling process, the data proceeds to a crucial pre-processing stage, where it undergoes transformations to enhance model comprehension.

The feature extraction process ensues, refining the dataset to improve its quality and informativeness. Subsequently, the dataset is partitioned into subsets for training and testing purposes. The subsequent phase involves the model classification process, utilizing various methods as outlined previously. This process generates predictions that undergo a rigorous evaluation to accurately gauge their performance. Evaluation metrics such as accuracy or precision are employed to ensure the effectiveness of the model in classifying multiple-choice questions.

Every step in this systematic approach contributes to the robustness and effectiveness of the multiple-choice question classification model. From data collection to performance evaluation, each stage is meticulously detailed to ensure the overall success and precision of this research in producing a reliable classification model.

3.2. Data Collecting

The data processed in this study are multiple-choice questions from elementary and junior high school levels. Multiple-choice questions are obtained from question banks, which are then collected according to school levels. The dataset consists of questions and difficulty labels. The construction of the dataset also involves manual labeling for the difficulty level categories of easy, medium, and difficult.

Table 1. Example of a question dataset

No.	Question	Label
1.	If the sum of angles in a triangle is 180 degrees, and two angles in the triangle are 50 degrees and 70 degrees, what is the measure of the remaining angle?	Medium
2.	A car moves at a constant speed of 60 km/h for 2 hours. How far does the car travel?	Easy
3.	Given a square has a side length of 5 cm. Determine the area and perimeter of the square.	Medium

Based on Table 1, there are sample questions, along with predefined labels. The labeling process was done manually and analyzed by teachers at each school level. The labels were assigned using a dataset of 413 questions. The labeling of the questions was divided into 3 level categories: easy, medium, and difficult. Each category was then assigned a value of -1 for the easy category, 0 for the medium category, and 1 for the difficult category.

3.3. Preprocessing

Data pre-processing is an important stage in system design that aims to improve the performance of algorithms while reducing their computational complexity [8]. This process also serves to prepare the text so that it is more effective and suitable for modeling needs. In this process, there are several steps taken. First, data cleaning is performed to remove double spaces, punctuation marks, and unwanted URLs. After that, typesetting is done to simplify the text. The next step is tokenization, which separates the text into words or small parts. Finally, normalization is used to simplify the text into a standard form. By going through this pre-processing process, the data becomes more structured and ready to be used in further analysis or model development.

3.4. Features

Existing data in the form of questions is processed through a data pre-processing stage before proceeding to the feature extraction process. In this final project, there are three main features used to predict scores. First, there is feature embedding, a concept in Natural Language Processing (NLP) where each word in a vocabulary is mapped into a numerical vector that represents the meaning of the word [9]. One of the embedding methods used is Word2Vec, which converts words into vectors so that the computer can understand the context of the meaning behind the words. Next, there are linguistic features, which include language markers that have special meanings and explain differences in the way languages are written. Linguistic features help in understanding the nuances of language and sentence structure that can affect judgment. Therefore, in this study, an in-depth understanding of linguistic features is used to recognize patterns in questions and make an important contribution in assessing the difficulty of such questions.

This feature extraction process is a key step in model development. Using methods such as embedding and linguistic analysis, researchers can convert text into a form that is understandable by computers. The result is the model's ability to predict scores based on the unique characteristics of each question. This process plays an important role in helping to improve understanding of the difficulty of questions and can aid in designing more effective and fair exams for participants.

3.4.1. Embeddings

In this research, an embedding extraction feature called Word2Vec is used. Word2Vec is a word vector representation created by Google. The reason for using Word2Vec is because this technology utilizes the Dense Matrix feature and is able to reduce the data dimension to be denser than the Term Frequency (TF) approach which uses Sparse Matrix. Therefore, it is expected that the use of Word2Vec can produce better results [9]. Word2Vec works by describing numerical vectors based on words that frequently co-occur in sentences. The Word2Vec model is trained to represent word vectors based on the dictionary and context of the frequently co-occurring words. The model is then able to extract feature vectors from the words used. The Word2Vec model used in this research has been trained using a corpus with a size of 100 vectors. The importance of using Word2Vec lies in the density of information it produces. This extraction feature allows extracting word vectors based on the co-occurrence of words in the text. For example, if two words frequently appear in the same context, Word2Vec can represent both words with semantically close numerical vectors. Using a pre-training model of the corpus, Word2Vec in this study consists of 100 vectors that are able to better describe the relationship between words, providing a solid foundation in analyzing and understanding text.

3.4.2. Linguistic

The linguistic features used in this research consist of two types, namely lexical features and syntactic features. Lexical features measure the density of grammar in a text. Lexical measurement based on text complexity calculates the total number of words with lexical properties divided by the total number of orthographic words [11]. The lexical feature used in this research is TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF method is a technique used to calculate the relative frequency of each word in a text [12]. This method assigns a value to each word based on the level of importance or the number of occurrences of the word in the document [12]. The TF-IDF equation can be seen in Equation 1.

$$TFidf = \frac{t}{d} \times idf \ dengan \ idf = log(\frac{N}{df(t)}) \tag{1}$$

Based on equation 1, where t is the number of occurrences of a word in document d, and d is the total words in the document. In the idf equation, N is the total number of documents, and df(t) is the number of documents containing word t. Syntactic features refer to features extracted from questions based on the structure of the words in the question [13]. The syntactic features used in this study are grammatical words. The resulting grammatical categories are generally nouns (NN), adjectives (JJ), and verbs (VERB). The application of syntactic feature analysis in this study uses the POS tagging method. Syntactic features apply POS tagging features for accurate word class labeling and provide grammatical information for each word in the sentence.

3.5. Text Classification

After the questions go through the feature extraction stage, the next step is the text modeling or classification process. Classification is the process of grouping documents into one or more predefined categories or into classes of similar documents [2]. In the field of Natural Language Processing (NLP), classification enables direct text analysis and category formation based on existing content. In this research, text classification methods are used to understand the characteristics of the questions based on the content contained in them.

The classification process is an important step in processing the questions that have gone through feature extraction. Classification makes it possible to map the questions into predefined categories. By using classification techniques, text analysis can be directly run to form groups based on the content of the question. In the context of this research, text classification is used to understand and categorize questions based on the information contained in the text.

The importance of the text classification process in this research lies in its ability to decipher the complexity of each question in greater depth. This process allows the researcher to identify unique patterns in the question text and categorize them according to their characteristics. Thus, the text classification method used in this research is an important foundation in understanding the essence of the questions that have gone through feature extraction.

3.5.1. Random Forest

One of the first methods implemented was Random Forest. The implementation of this model uses a library in the Python programming language. The output of this model is the prediction of question difficulty based on the dataset that has gone through the pre-processing stage of data feature extraction. The Random Forest algorithm used is one of the well-known machine learning methods, where various decision trees are inserted into each part of the dataset, then the prediction results are obtained using the average value of the decision tree results [14]. In the decision tree structure, there are root nodes, internal nodes, and leaf nodes, which play an important role in the prediction process [15].

The implementation stage of the Random Forest method involves several steps. First, the number of trees (k) is determined from all features (m), where k is a smaller number than the total number of features. Next, random sampling (N) of the dataset is performed for each tree. After that, a random subset is drawn, consisting of m predictors, where m is a smaller number than the total number of predictor variables (p). The second and third steps are repeated k times until reaching the specified number of trees [16]. Using this method, the system is able to provide question difficulty prediction by effectively utilizing decision tree clustering.

3.5.2. Support Vector Machine

The second method used is SVM. SVM is a classification algorithm that functions to classify nonlinear data and linear data [17]. The SVM concept is used to find the best hyperplane, which is very important for delimiting two classes. The best hyperplane can be obtained by measuring the margin of the hyperplane and determining the best point. In SVM, it is important to find the optimal decision boundary, which is the hyperplane with the largest margin between two classes. This margin is measured as the distance between the hyperplane and the closest point of each class. The selection of the hyperplane with the largest margin will result in a more accurate classification model. In addition, SVM can also handle nonlinear data by using a kernel transformation technique, which transforms the data to a higher dimension so that it can be separated by the hyperplane. Thus, SVM is a strong choice to overcome the challenges of complex data classification and ensure optimal class separation.

3.5.3. Gaussian Process

The third method used is the Gaussian Process. Gaussian Process is a non-parametric process that can naturally generate probabilities [18]. The model used is Naïve Bayes, which uses Bayes' Theorem. Naïve Bayes assumes the value of an attribute in a class is independent of other values. Based on equation 2, X is evidence, H is hypothesis, P(H|X) is the posterior probability of H conditional on X, P(X|H) is the posterior probability of X conditional on H, P(H) is the prior probability of hypothesis H, P(X) is the prior probability of evidence X. This method is used to identify the relationship between evidence and hypothesis, taking into account the prior probability of the hypothesis as well as the evidence, thus helping in determining the posterior probability of a hypothesis based on existing evidence. Naïve Bayes has simple assumptions but often gives good results in data classification by combining information from various attributes efficiently. The application of Naïve Bayes helps in understanding data patterns by considering probabilities and relationships between attributes. In this context, the method is used as a tool to measure the level of confidence in hypotheses based on existing evidence, facilitating intelligent and informed decision-making.

$$P(X) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

3.5.4. Dense Neural Network

The third method used is the Gaussian Process. Gaussian Process is a non-parametric process that can naturally generate probabilities [18]. The model used is Naïve Bayes, which uses Bayes' Theorem. Naïve Bayes assumes the value of an attribute in a class is independent of other values. Based on equation 2, X is evidence, H is hypothesis, P(H|X) is the posterior probability of H conditional on X, P(X|H) is the posterior probability of X conditional on H, P(H) is the prior probability of hypothesis H, P(X) is the prior probability of evidence X. This method is used to identify the relationship between evidence and hypothesis, taking into account the prior probability of the hypothesis as well as the evidence, thus helping in determining the posterior probability of a hypothesis based on existing evidence. Naïve Bayes has simple assumptions but often gives good results in data classification by combining information from various attributes efficiently. The application of Naïve Bayes helps in understanding data patterns by considering probabilities and relationships between attributes. In this context, the method is used as a tool to measure the level of confidence in hypotheses based on existing evidence, facilitating intelligent and informed decision-making.

$$\begin{aligned} A_x = [a_{11}\, a_{12} \cdots \ a_{12}\, a_{21}\, a_{22} \cdots \ a_{2n} \ \vdots \ \vdots \ \ddots \\ \vdots \ a_{m1}\, a_{m2} \cdots \ a_{mn}][x_1\, x_2 \ \vdots \ x_n] \\ = [a_{11}x_1 + \ a_{12}x_2 + \cdots + \ a_{1n}x_n\, a_{21}x_1 + \ a_{22}x_2 + \cdots \\ + \ a_{2n}x_n \ \vdots \ \vdots \ \ddots \\ \vdots \ a_{m1}x_1 + \ a_{m2}x_2 + \cdots + \ a_{mn}x_n] \end{aligned} \tag{3}$$

3.5.5. Logistic Regression

The fifth method used is Logistic Regression. Logistic regression uses probability to predict a classification [19]. Logistic regression aims to identify the relationship between independent variables and one or more dependent variables by using probability as a predictive value for the dependent variable. The logistic regression formula can be seen in equation 4. Based on equation 4, P is the probability based on the value of the independent variable. e is the Euler constant used to calculate the exponential function. β_n is the coefficient or weight determined during the model training process. x1, x2, ... are the values of the independent variables used to predict the probability of the target variable y. In this context, logistic regression is used to estimate the success or failure category of an event based on predefined independent variables. Logistic regression is one of the important tools in predictive analysis and is used to understand the probability of events occurring in a dataset. The use of logistic regression in this context is to assess the odds of an outcome based on the variables involved, and the model can adapt itself to the given data to provide accurate predictions regarding the odds of success or failure. Thus, logistic regression plays a crucial role in understanding the correlation between the variables involved and evaluating the probability of success of an event.

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{4}$$

3.6. Multiclass Confusion Matrix

After the modeling process, the last step is the evaluation process. This process will measure the performance and classification results with a matrix. The Confusion Matrix is very commonly used in machine learning for supervised classification or determining classification models [20]. The confusion matrix analyzes the extent to which a classification model recognizes various classes [16]. This study uses three categories to implement a multiclass confusion matrix with NxN dimensions to cover all possible combinations of predicted and true classes [21]. The output results include accuracy, recall, precision, and F1-Score based on the question dataset that has passed the classification process with five methods.

$$Accuracy = \frac{(TP1+TP2+\cdots+TPn)}{(TP1 + TP2 + \ldots + TPn + FP1 + FP2 + \ldots + FPn + FN1 + FN2 + \ldots + FNn)} \tag{5}$$

$$Recall = \frac{(TP1+TP2+\cdots+TPn)}{(TP1 + TP2 + \ldots + TPn + FN1 + FN2 + \ldots + FNn)} \tag{6}$$

$$Precision = \frac{(TP1 + TP2 + \ldots + TPn)}{(TP1 + TP2 + \ldots + TPn + FP1 + FP2 + \ldots + FPn)} \tag{7}$$

$$F1 - Score = 2 \times \frac{(Presisi \times Recall)}{(Presisi + Recall)} \tag{8}$$

Based on equation 5, accuracy calculates all correct predictions divided by the number of queries. Equation 6 shows recall, which is the number of actual positive questions that were correctly predicted. Equation 7 is precision, which is the number of questions or queries predicted correctly. Meanwhile, equation 8 is the F1-score, which is the harmonic mean between recall and precision. In this context, TP is True Positive, FP is False Positive, and FN is False Negative. TP indicates when the prediction result of the question matches the actual condition, while FP and FN indicate when the prediction result does not match the actual condition. Using such information, the evaluation of the model's performance in identifying correct questions can be measured using these metrics, which provide a comprehensive picture of how well the model can predict and classify the given questions.

Discussion: The proposed model is tested on different parameters, and it was observed that dividing the image into a grid gives better results than using the original image. Further, tests were performed to check the best value of k (number of neighbors in the nearest neighbor algorithm). It was found that k=9 gives the best results.

4. Result and Discussion

This study predicts question difficulty using a dataset of multiple-choice questions at primary and secondary school levels labeled as easy, medium, and difficult. Three types of feature extraction are used, namely embedding, lexical features, and syntactic features. Then this research continues by applying five classification methods: Random Forest, Support Vector Machine, Gaussian Process, Dense Neural Network, and Logistic Regression. The question difficulty prediction classification was built with five scenarios. The first scenario determines the classification prediction results of each method using embedding feature extraction. The second scenario determines the classification prediction results of each method using lexical feature extraction. The third scenario determines the classification prediction results of each method using syntactic feature extraction. The fourth scenario determines the classification prediction results of each method using lexical feature extraction and syntactic features. The fifth scenario determines the classification prediction results of each method using embedding feature extraction and syntactic features. The last scenario determines the prediction results with a dense neural network model without involving feature extraction. Then the classification results of each scenario are compared and analyzed.

4.1. Data

In this study, 413 questions were used. This data uses three class categories, namely easy, medium, and difficult. The distribution of the labeled data obtained can be seen in Figure 3.

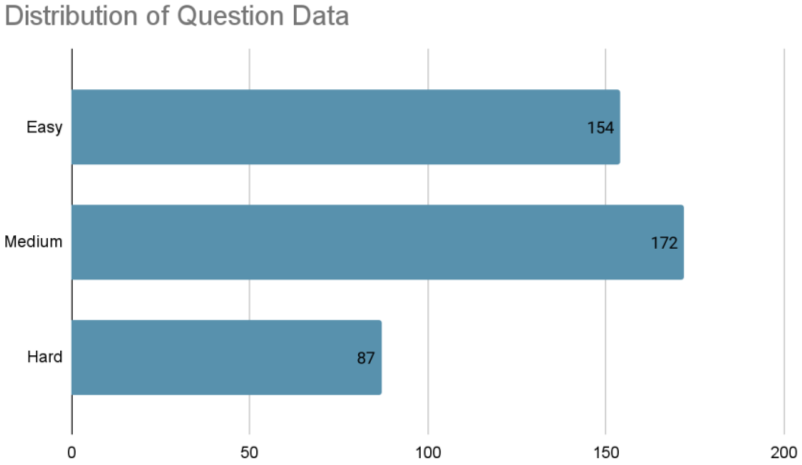

Figure 3. Distribution of question labels

4.2. Classification with Embedding Features

The purpose of scenario 1 is to compare the best accuracy results of the classification models used. In the first scenario, the accuracy of the model will be tested by applying embedding feature extraction to the dataset. The embedding feature used is Word2Vec, which converts words into vectors. The results of the first scenario can be seen in Table 4.

Table 4. Results of accuracy classification using feature embedding

Model	**School**	
Ratio (20:80)	**Elementary**	**Junior-School**
Random Forest	0,71	0,80
SVM	0,57	0,76
Naïvel Bayes	0,55	0,59
Logistic Regression	0,71	0,73

Based on Table 4, the Random Forest classification model achieved the highest accuracy results. This experiment used a 20:80 split ratio with 20 test data and 80 training data. The accuracy in Random Forest using embedding feature extraction resulted in relatively high values in all trials of elementary and junior high school question datasets. The highest accuracy for the first scenario was 80%. Meanwhile, the lowest accuracy was obtained by SVM and Naïve Bayes models with 57% and 55% accuracy, respectively. This happens because SVM uses the optimal linear separator hyperplane so that SVM is more complex in linear mapping on small datasets. Meanwhile, naïve bayes considers the features in the dataset to be independent of each other.

4.3. Classification with Lexical Features

The purpose of the second scenario is to compare the best accuracy results of the classification model using lexical feature extraction. The lexical feature used is TF-IDF. TF-IDF feature extraction will give value to words that often appear in sentences. The results of the second scenario can be seen in Table 5.

Table 5. Results of accuracy classification using lexical features

Model	**School**	
Ratio (20:80)	**Elementary**	**Junior-School**
Random Forest	0,71	0,84
SVM	0,60	0,80
Naïve Bayes	0,57	0,67
Logistic Regression	0,65	0,80

Based on Table 5, the Random Forest model obtained the highest accuracy results. This second scenario also uses a 20:80 split ratio with 20 test data and 80 training data. TF-IDF feature extraction produces relatively high accuracy in the Random Forest classification model. The highest accuracy in the second scenario was 84%. This is because the Random Forest model uses a collectively decision tree. The decision tree looks at a subset of words at each iteration, which means that TF-IDF will be more informative in weighting different words.

4.4. Classification with Syntactic Features

The purpose of the third scenario is to compare the best accuracy results of the classification model using syntactic feature extraction. The syntactic feature used is the POS (Part-of-Speech) identifier. POS identifiers will pay attention to the grammatical structure of each word in the sentence by labeling each word based on the type of word. The results of the third scenario can be seen in Table 6.

Table 6. Results of accuracy classification using syntactic features

Model	**School**

Ratio (20:80)	**Elementary**	**Junior-School**
Random Forest	0,80	0,76
SVM	0,36	0,63
Naïve Bayes	0,68	0,42
Logistic Regression	0,27	0,60

Based on Table 6, the Random Forest model produces the highest accuracy results. POS tagging feature extraction produces relatively high accuracy in the Random Forest and SVM classification models. The highest accuracy in the third scenario is 80%. POS tag features can provide information representation to the Random Forest model to utilize word context.

4.5. Classification with DNN Model

The sixth scenario compares the best accuracy results of the Deep Neural Network (DNN) classification model. DNN uses each neuron to receive input from the previous layer and provides output to all neurons in the next layer. The results of the fourth scenario can be seen in Table 7.

Table 7. Results of accuracy classification using feature syntactic

Model	**School**	
Ratio (20:80)	**Elementary**	**Junior-School**
DNN	0,71	0,80

Based on Table 7, the highest accuracy is obtained on a dataset of 80%. In this scenario, only direct modeling with the Dense Artificial Neural Network is used. This scenario does not involve any feature extraction, as is the case with the other scenarios.

Table 8. Comparison of the accuracy levels of each model

Feature Extraction	**Elementary**				**Junior-School**			
	RF	**SVM**	**Naïve Bayes**	**Logreg**	**RF**	**SVM**	**Naïve Bayes**	**Logreg**
Embedding	0,71	0,55	0,57	0,71	0,80	0,76	0,59	0,73
TF-IDF	0,71	0,60	0,57	0,65	0,84	0,80	0,67	0,80
Syntactic Features	0,80	0,36	0,68	0,72	0,76	0,63	0,42	0,60
Model: DNN	0,71				0,80			

Based on Table 8, we can see the accuracy comparison of each model with different scenarios and from different datasets. Starting from the question dataset, we get the highest accuracy result of 84% using the Random Forest model. This accuracy result is obtained through tf-idf feature extraction before modeling. If we analyze the overall accuracy, Random Forest is the best model with high accuracy in predicting question difficulty. In addition, the best feature extraction is Embedding and TF-IDF. Of course, all the accuracy that has been obtained in each model involves data pre-processing, which is very influential in optimizing the accuracy results. The data used in the study also has an influence, especially on the label balance. If we look at the dataset used, there is an imbalance between the three labels, with more labels on easy questions. This can affect the training data trained on the classification model and also affect the prediction results. The experimental results show that the best classification method in identifying question difficulty is Random Forest, which is similar to the findings of a previous study by Le An Ha [3]. In the best feature extraction, Le An Ha came up with embedding as the best feature extraction, and in this study, embedding and TF-IDF became the best feature extraction.

5. Conclusion

In this study, we conducted a classification analysis to predict the difficulty levels of multiple-choice questions at primary and secondary school levels, employing five methods—Random Forest, SVM, Naïve Bayes, Logistic Regression, and DNN—and three feature extractions—embedding, lexical, and syntactic. The results demonstrated that the Random Forest method consistently outperformed others, achieving the highest accuracy in each question type across various scenarios. The comparison table reinforced the superiority of the Random Forest method, indicating its reliability in predicting question difficulty. Additionally, the choice of feature extraction played a significant role, with embedding and TF-IDF emerging as the most effective methods. Despite these promising findings, it is crucial to acknowledge certain limitations in this research, notably the potential for data imbalance due to the unequal distribution of questions across difficulty levels. Furthermore, the study focused on questions from only two subjects, limiting its generalizability. To address these limitations, future research should strive for balanced datasets, encompassing a broader range of subjects to enhance the model's classification capabilities.

This study contributes valuable insights to the field of educational assessment by demonstrating the effectiveness of machine learning methods in predicting the difficulty levels of multiple-choice questions. The identified impact lies in the practical application of the Random Forest method, which can aid educators in crafting assessments tailored to students' comprehension levels. The emphasis on feature extraction, particularly the success of embedding and TF-IDF, adds a nuanced layer to the understanding of how these techniques influence prediction accuracy. The research impact extends to the broader educational landscape, where the integration of machine learning can potentially streamline the assessment process, providing more nuanced and personalized feedback to students and educators alike. As technology continues to play an increasing role in education, the findings of this study offer a foundation for future developments in adaptive assessment systems and data-driven educational practices.

6. Declarations

6.1. Author Contributions

Conceptualization: P.S.S. and R.G.H.; Methodology: B.H.H; Software: P.S.S.; Validation: P.S.S. and R.G.H.; Formal Analysis: P.S.S. and B.H.H.; Investigation: B.H.H.; Resources: P.S.S.; Data Curation: R.G.H.; Writing Original Draft Preparation: P.S.S. and B.H.H.; Writing Review and Editing: R.G.H; Visualization: B.H.H.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

References

[1] A. Das, "An alternative approach for question answering system in Bengali language using classification techniques," *INFOCOMP J. Comput. Sci.*, vol. 19, no. 1, pp. 1-8, 2020.

[2] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nat. Commun.*, vol. 10, no. 1, pp. 3096-3104, 2019.

[3] T. P. Sahu, R. S. Thummalapudi, and N. K. Nagwani, "Automatic question tagging using multi-label classification in community question answering sites," in *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, IEEE, vol. 1, no.1, pp. 63-68, 2019.

[4] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, no. 1, pp. 325-338, 2019.

[5] O. M. Crook, T. Smith, M. Elzek, and K. S. Lilley, "Moving spatial profiling proteomics beyond discrete classification," *Proteomics*, vol. 20, no. 23, pp. 1900392-1900400, 2020.

[6] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, "Retrieving and reading: A comprehensive survey on open-domain question answering," *arXiv Prepr. arXiv2101.00774*, vol. 1, no. 1, pp. 1-12, 2021.

[7] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv Prepr. arXiv2001.07676*, vol. 1, no. 1, pp. 1-13, 2020.

[8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: a comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1-40, 2021.

[9] K. Makhlouf, L. Amouri, N. Chaabane, and E.-H. Nahla, "Exam Questions Classification Based on Bloom's Taxonomy: Approaches and Techniques," in *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, IEEE, vol. 1, no. 1, pp. 1-6, 2020.

[10] B. T. Hung, "Integrating diacritics restoration and question classification into vietnamese question answering system," *Adv. Sci. Technol. Eng. Syst. J*, vol. 4, no. 1, pp. 207-212, 2019.

[11] S. Chakraborty, S. Paul, and M. Rahat-uz-Zaman, "Prediction of apple leaf diseases using multiclass support vector machine," in *2021 2Nd international conference on robotics, electrical and signal processing techniques (ICREST)*, IEEE, vol. 1, no.1, pp. 147-151, 2021.

[12] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, no. 1, pp. 90847-90861, 2020.

[13] R. Sarki, K. Ahmed, H. Wang, Y. Zhang, and K. Wang, "Convolutional neural network for multi-class classification of diabetic eye disease," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 9, no. 4, pp. 1-13, 2021.

[14] R. L. Kumar, J. Kakarla, B. V. Isunuri, and M. Singh, "Multi-class brain tumor classification using residual network and global average pooling," *Multimed. Tools Appl.*, vol. 80, no. 1, pp. 13429-13438, 2021.

[15] R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, "Automated detection of mild and multi-class diabetic eye diseases using deep learning," *Heal. Inf. Sci. Syst.*, vol. 8, no. 1, pp. 32-40, 2020.

[16] S. D. A. Bujang *et al.*, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, no. 1, pp. 95608-95621, 2021.

[17] K. Purwandari, J. W. C. Sigalingging, T. W. Cenggoro, and B. Pardamean, "Multi-class weather forecasting from twitter using machine learning aprroaches," *Procedia Comput. Sci.*, vol. 179, no. 1, pp. 47-54, 2021.

[18] M. Koklu and I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," *Comput. Electron. Agric.*, vol. 174, no. 1, pp. 105507-105514, 2020.

[19] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Min.*, vol. 14, no. 1, pp. 1-22, 2021.

[20] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, no. 1, pp. 216-231, 2019.

[21] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets-a first step towards preventing skin cancer," *Neurosci. Informatics*, vol. 2, no. 4, pp. 100034-100045, 2022.

[22] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-label confusion matrix," *IEEE Access*, vol. 10, no. 1, pp. 19083-19095, 2022.

[23] D. Suhartono dan K. Khodirun, "System of Information Feedback on Archive Using Term Frequency-Inverse Document Frequency and Vector Space Model Methods," *International Journal of Informatics and Information Systems,* vol. 3, no. 1, pp. 36-42, 2020. doi: 10.47738/ijiis.v3i1.6.

[24] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv Prepr. arXiv2008.05756,* vol. 1, no. 1, pp. 1-13, 2020.

[25] K. Fujishima, “A Study on Hospitality Education at University : Jal’s Philosophy Education as an Example”, *Int. J. Appl. Inf. Manag.,* vol. 1, no. 3, pp. 136–144, Jul. 2021.

[26] I. Markoulidakis, G. Kopsiaftis, I. Rallis, and I. Georgoulas, "Multi-class confusion matrix reduction method and its application on net promoter score classification problem," in *The 14th pervasive technologies related to assistive environments conference*, vol. 1, no. 1, pp. 412-419, 2021.