POGIL Data Analysis Employing Classification Algorithms to Examine Student Performance

Sahil Parab, Priyanshu Singh, Anjali Yeole, Maya Bhat

Artificial Intelligence & Data Science, VESIT, Mumbai, India

Corresponding author: Sahil Parab, Email: 2020.sahil.parab@ves.ac.in

Education is a fundamental and essential necessity that enhances the potential of everyone. It holds significance for achieving life goals and personal development. This study employs data mining classification techniques to analyze student data from the Vivekananda Education Society Institute of Technology in Chembur, Mumbai. The goal is to determine whether there exists a discernible pattern between the grades achieved by students instructed using POGIL (Process Oriented Guided Inquiry Learning) and those taught through traditional methods. POGIL represents a collaborative learning approach that integrates Guided Inquiry into a cyclic framework of concept generation, investigation, and application [2].

The student cohort was divided into two groups: one received instruction through the POGIL methodology, while the other experienced traditional teaching methods. Subsequent tests were administered to assess the performance of each group. Upon analyzing the collected data, it was revealed that students exposed to the POGIL learning method exhibited superior scores. Finally, a predictive model was developed to estimate the potential score increase for students adopting the POGIL methodology.

Keywords: J48 decision tree, K-nearest neighbour, Random Forest Classifier, logistic Regression, WEKA.

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 569–581. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-52

1. Introduction

Student-centered, group-learning instructional technique and philosophy, known as Process Oriented Guided Inquiry Learning (POGIL), was created as a result of research into the most powerful strategies for learners to study. POGIL was developed in 1994 to improve general chemistry instruction. Today, more than 2000 schools and colleges follow the POGIL technique in their teaching.

The design of a POGIL exercise must prioritize these two factors [2]. Firstly, it aims for students to generate the desired notions during the initial "Exploration" phase. The guiding questions must also be properly prepared to enable students not only to arrive at the correct conclusions but also to acquire a variety of process and learning abilities. The initial questions usually build upon students' existing knowledge and draw attention to the details that the model has to offer. Subsequent questions are meant to encourage the identification of correlations and patterns in the data, and eventually, some concept development follows. The last few questions may ask students to generalize their newly acquired information and understanding and apply the ideas to fresh circumstances. POGIL activities thus adhere to the learning cycle pattern of concept invention, exploration, and applications. Constructiveness serves as a solid foundation for POGIL instruction [1][13][14].

The purpose of analyzing student performance is to ascertain if there are any patterns between values and relationships among data entities where students learned with POGIL. Many algorithms, including Naive Bayes, J48 decision trees, K-nearest neighbor, and others, are used in data mining to analyses student performance and extract knowledge, such as classification. This information can help forecast student progress by revealing significant correlations and unexpected outcomes [1]. Fig 1 represents the process skills to be included in the students. Due to POGIL techniques students are developing like information processing, criticalthinking, problemsolving.

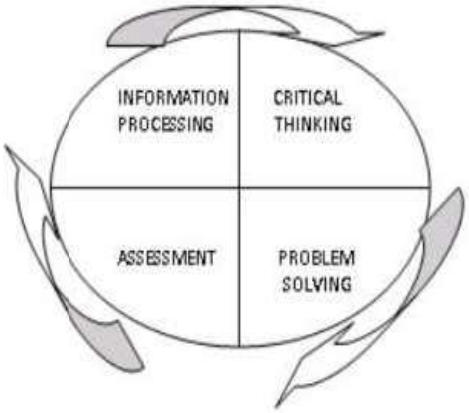

Figure 1: Process skills to be included in students

2. About Student Dataset

We have two datasets: one with POGIL and one without POGIL. The dataset with POGIL includes attributes such as JEE, SSC, HSC, Admittance, Final, and Merit, consisting of 118 sets of student details Fig 3. In total, the dataset encompasses information for 545 students, with 7 columns containing JEE, SSC, HSC, Admittance, Final, and Merit attributes.

	SSC	Chemistry	HSC	Merit	JEE	Final	Admittance
count	545.000000	545.000000	545.000000	545.000000	545.000000	545.000000	545.000000
mean	88.086991	71.181651	79.237083	3080.705085	69.359633	34.858716	0.508257
std	6.542388	13.585546	8.874950	21577.553746	31.519996	9.839159	0.500391
min	60.000000	38.000000	54.400000	19.636120	-10.000000	0.000000	0.000000
25%	85.090000	61.000000	73.080000	72.485951	46.000000	28.000000	0.000000
50%	89.640000	71.000000	80.310000	85.738836	68.000000	36.000000	1.000000
75%	92.910000	82.000000	86.460000	93.378181	89.000000	41.000000	1.000000
max	99.270000	98.000000	97.500000	383855.000000	178.000000	59.000000	1.000000

Figure 2: Summary of Data Set

By analyzing the correlation heat map in Figure 2, we can effectively pinpoint the highly correlated attributes within the dataset. Specifically, a robust correlation is noticeable between the 'Final' and 'Admittance' attributes. Additionally, we can discern relationships between 'HSC,' 'Chemistry,' and 'SSC' within the dataset through this correlation analysis.

Figure 3: Confusion Matrix

3. Data Mining

The process of data mining, referred to as knowledge discovery in data (KDD), is utilized to extract patterns and other significant information from vast datasets [10]. Gartner Inc. provides one of the most comprehensive definitions of data mining, describing it as "the process of discovering new patterns, trends, and meaningful correlations between attributes from data stored in company databases, achieved through statistical and mathematical techniques."

Smart data analytic have greatly enhanced corporate decision-making through data mining. The data mining methods underpinning these analyses can be categorized into two groups: those characterizing the target dataset and those predicting outcomes using machine learning algorithms. These approaches are instrumental in surfacing valuable insights, including fraud detection, user habits, identifying bottlenecks, and even detecting security breaches, through data organization and filtering.

In an emerging field known as "educational data mining," methods are being developed to examine the distinctive and continually expanding volumes of data generated in educational settings. The aim is to utilize these methods to gain deeper insights into students and the learning environments in which they are taught [7].

4. Data Mining Process

Demand for reliable and standardized data mining approaches is rising quickly. Cross-Industry Standard Method for Data Mining is the most used one (CRISP-DM) [10]. The reliable data mining paradigm CRISP-DM contains six stages. This cyclical strategy gives the data mining process a structured approach. Although the six phases can be carried out in any order, there may be times when it's necessary to go back and repeat steps. Business Knowledge: The firm's goals are established at this stage, and important components that will help achieve the goal are recognized. collecting data and populating tool data the data will be gathered at this point (if using any tool). The data is presented along with the data source, location, acquisition method, and any issues that might have occurred. To guarantee the accuracy of the data, it is visually inspected and queried. Selecting the appropriate data, cleaning it, constructing attributes from it, and combining data from many sources are all parts of data preparation[4]. Modelling includes selecting a data mining method, like decision trees, coming up with a test design to gauge the model's performance, building models from the dataset, and debating the findings with subject-matter experts. Evaluation: This process will determine how well the finished model satisfies the business requirements. By putting the model through applications in the actual world, it may be assessed. The model is reviewed to check for mistakes or steps that should be repeated. Deployment: At this phase, a deployment plan is made, a strategy to monitor and maintain the results of the data mining model to evaluate their usefulness is developed, final reports are prepared, and the entire process is examined to check for faults and to see if any steps need to be repeated [2]-[5].

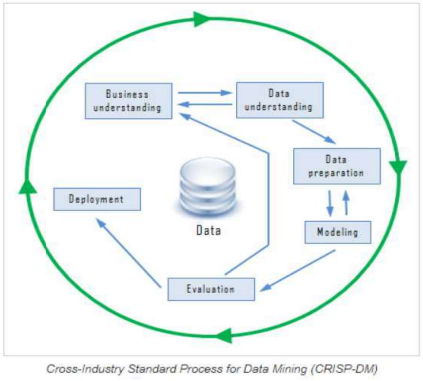

Figure 4: Data mining chart.

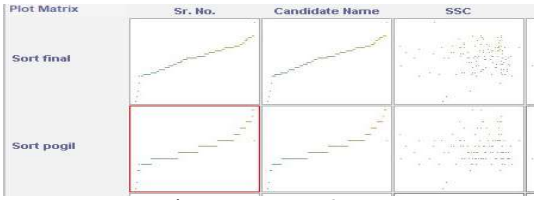

Figure 5: Data Analysis.

Waikato Environment for Knowledge Analysis (WEKA)
The WEKA software is the most popular and widely used machine learning software in the industry. It is written in Java and was developed by the University of Waikato in New Zealand. Many algorithms and big data sets are supported by WEKA. It is open-source software that is distributed in accordance with the GNU General Public License. It includes various tools for data pre-processing and provides a bunch of algorithms like clustering, association rules, regression, and classification. Primarily used for predictive modelling and data analysis. Its graphical user interface makes accessing its numerous functionalities simple. [3]

WEKA version 3.8.1 was used for its open-source nature, portability since it was developed using Java and hence runs on most computing platforms, its comprehensive tools, and ease of use. [3]

Here we have uploaded a csv dataset file to WEKA for analysis. It supports many formats of files. Figure 6 displays the statistics concerning the students who utilized the POGIL technique for learning. From the data presented in the figure, it is evident that 68.50% of the students achieved good scores, while 31.50% of the students attained poor scores. In contrast, Figure 7 portrays the statistics of students who did not employ the POGIL technique. The notable disparity in scores is evident, with 49.57% of students achieving good marks and 50.43% of students attaining poor scores.

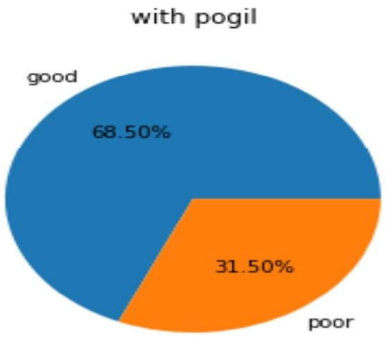

Figure 6: Results after using POGIL technique.

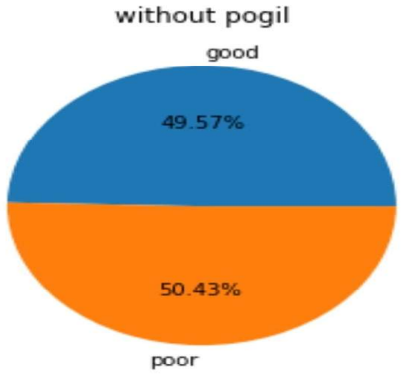

Figure 7: Result without using POGIL technique.

4.1 Classification

Data is categorized using the process of classification into a predetermined number of classes. It is a supervised learning method that requires a labelled dataset to train and build the model. The classification model classifies to which class the input values belong and to which class the model is trained on the data. It will predict the class categories for the new data. A feature is a specific, quantifiable characteristic of the phenomenon under observation.

Figure 8: Classification process

Classification algorithms could be broadly classified as linear classifiers, support vector machines, quadratic classifiers, kernel estimation, decision trees, neural networks, and learning vector quantization.

The approach is based on machine learning. A data set's objects are categorized using the classification process into a predetermined number of groups. During the training, the machine learns the weights of attributes and classifies the output.[3]

4.2 C4.5 algorithm/J48

A classification algorithm called C4.5 generates information-theoretic decision trees. It is an improvement on Ross Quinlan's earlier ID3 technique, which is based on the Java programming language and is frequently referred to in Weka as J48. C4.5 is sometimes referred to as a statistical classifier because the decision trees it produces are used for categorization. Numerous additional features, such as accounting for missing data, decision tree pruning, continuous attribute value ranges, rule derivation, etc., are included in the J48 implementation of the C4.5 method. The Weka data mining program uses the C4.5 algorithm, which is open-source and implemented in Java by J48. J48 allows classification using either decision trees or rules built from them.

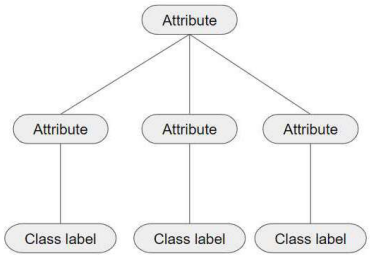

Figure 9: Decision Tree.

Similar to the ID3 algorithm, this approach shares the objective of constructing decision trees based on a set of training data. This method hinges on the concept of information entropy, a measure of uncertainty or disorder within a dataset. The training data is comprised of a collection of previously classified samples, often represented as S = {s1, s2, ...}. Each individual sample, denoted as si, is characterized by a p-dimensional vector (x1i, x2i, ..., xpi), where each xj signifies the values of various attributes or features associated with the sample, along with the class to which it belongs.

The critical decision in constructing these decision trees is determining which attribute to use for splitting the data at each node. The selection is based on maximizing classification accuracy. In essence, the algorithm evaluates the information gain or entropy reduction that each attribute provides. The attribute that offers the most information, often measured as the reduction in uncertainty, is chosen as the dividing criterion.

This process continues recursively as the algorithm builds the decision tree, with each node representing a decision point based on an attribute. The goal is to organize the data in a hierarchical structure that facilitates effective classification, ultimately leading to more accurate predictions or categorizations based on the attributes of the samples.[3]

The C4.5 algorithm selects the property of the data from each node of the tree that divides its set of samples into subsets, enriched in one class or the other, in the most efficient manner. The difference in entropy is used to determine the splitting criteria, which is the normalized information gain. The attribute with the greatest normalized information gain is the one that is used to make the decision. Then, using a divide-and-conquer strategy, the C4.5 algorithm recurses on the partitioned sub lists and generates a decision tree based on the greedy algorithm.

4.2.1 Logistic Regression

Logistic regression is a supervised learning machine model. Logistic regression can be used for regression and classification problems of machine learning. It can be used to show the relationship between dependent and independent variables. The primary purpose of classification is to compare the difference between accepted and not accepted students based on their performance. It tells the total percentage of students whose performance is improved after using the POGIL technique [6]. Logistic regression uses the function called the sigmoid function figure 9. It has an 'S-shaped graph. The minimum and maximum value of the function is 0 and 1. All the dataset is normalized and fitted into this S-shaped function. We are passing the data into the model and predicting the admittance of students based on their performance. The model will predict weather the student is admitted or not based on the performance.

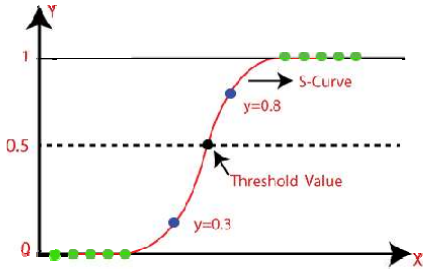

Figure 10: Logistic Function [9].

Result summary:

Correctly classified Instance	25
Kappa statistic	1
Correlation coefficient	0.9499
Mean absolute error	0.1192

Root mean squared error	0.1192
Relative absolute error	26.7179 %
Root relative squared error	24.6878 %
Total Number of Instances	25

Figure 11: Result of Logistic Regression.

In evaluating the model's performance figure 10, we observe that 25 instances have been accurately classified, indicating a strong predictive capability. The Kappa statistic, with a value of 1, further reinforces the model's accuracy and reliability, affirming a high level of agreement between the observed and predicted values. Additionally, the correlation coefficient stands at 0.9499, denoting a robust positive correlation between the predicted and actual outcomes. Mean absolute error and root mean squared error both measure at 0.1192, indicating a relatively low level of error in the model's predictions. In terms of relative error metrics, the relative absolute error is calculated at 26.7179%, and the root relative squared error stands at 24.6878%, providing insights into the error percentages relative to the actual values. Overall, these metrics, combined with a total of 25 instances, offer a comprehensive evaluation of the model's performance and accuracy.

4.2.2 Random Forest Tree

Random forest classifier is a supervised machine learning technique. In machine learning, it can be applied to both classification and regression problems. It is mainly used for classification problems in machine learning. It uses the concept of ensemble learning which means it combines the results of multiple decision trees based on majority votes to predict the values [6]-[7].

It takes a different training dataset and each training dataset has a decision tree associated with it. similarly, there is a combination of multiple decision trees we take the majority voted value as the result. As we are taking the student dataset for the analysis of the POGIL technique. We are going to predict the admittance of the student based on the performance of the student.

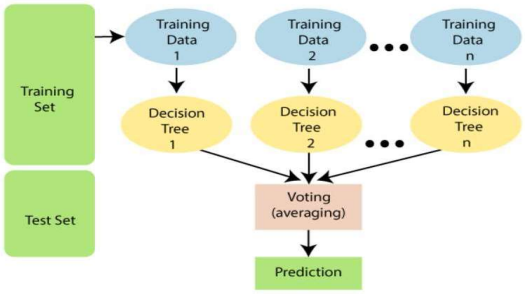

Figure 12: Random Forest Tree.

Random forest combines several different decision trees to predict the dataset's class, it is possible that some decision trees will predict the right outcome while others will not. But when viewed collectively, every tree forecasts the right result. Consequently, there are two presumptions for a better Random Forest classifier:

There should be some genuine values in the feature variable of the dataset for the classifier to forecast accurate outcomes as opposed to an assumed outcome.

The forecasts from each tree must have very little association with one another.

Result Summary:

Correlation coefficient	0.9499
Mean absolute error	0.0791
Root mean squared error	0.1076
Relative absolute error	50.8611 %
Root relative squared error	55.8827 %
Total Number of Instances	127

Figure 13: Result analysis of Random Forest.

The analysis of the model's performance reveals significant insights figure 12. The correlation coefficient, calculated at 0.9499, demonstrates a strong positive correlation between predicted values and actual outcomes, indicating a high level of accuracy in the predictions. Furthermore, the mean absolute error is computed to be 0.0791, showcasing a relatively low level of average error in the predictions made by the model. Similarly, the root mean squared error is evaluated at 0.1076, reaffirming the model's precision in predicting values closely to the actual data points. On the other hand, the relative absolute error and root relative squared error are expressed as percentages, highlighting the error rates relative to the actual values. The relative absolute error stands at 50.8611%, and the root relative squared error is 55.8827%, providing important insights into the model's performance from a relative error perspective. This analysis is based on a dataset comprising a total of 127 instances, contributing to a comprehensive assessment of the model's predictive capabilities.

4.2.3 K Nearest Neighbor

K Nearest Neighbor is a simple machine-learning algorithm for supervised learning. It can be used for regression and classification problems in machine learning [7]. It assumes the similarity between new cases and available cases and puts the data into the most similar category. It creates a category of similar types of objects, and when new data points come in, it checks the similarity between them and puts them in a special category. It is also known as the "lazy learner algorithm" since it classifies the dataset at prediction time rather than learning during training[6]. As we have a student dataset, we are classifying them into two categories, whether they are accepted or not. As we have a student dataset, we are classifying them into two categories, whether they are accepted or not. We are providing the input attribute as Chemistry, HSC, SSC, etc. to the model, and it is classified into two categories. The main purpose of classification is to compare the differences between accepted and non-accepted students based on their performance. It tells the total percentage of students whose performance improved after using the POGIL technique.

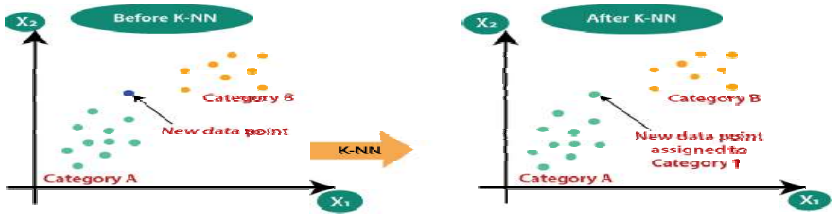

Figure 14: KNN Clustering[9].

The K-Nearest Neighbours' (KNN) algorithm operates by identifying the K closest neighbouring data points (instances) to a given point within the feature space, employing a similarity metric such as the Euclidean distance. The value of K is a pivotal parameter, influencing the number of neighbouring points considered during classification. The KNN process involves several key steps. Initially, a value for K is chosen, representing the quantity of nearest neighbours to consider. Subsequently, distances between the point to be classified and all other points in the dataset are calculated, often using the Euclidean distance formula. The K nearest neighbours to the target point are then determined based on the calculated distances. For classification tasks, a majority vote is conducted among these K neighbours to ascertain the prevalent class. The new point is assigned the class that appears most frequently among its K nearest neighbours. In a visual representation, like a figure 13, the data points are typically displayed in different classes, for instance, denoted by different colours (e.g., red points and blue points). The KNN algorithm utilizes this information to classify a new point by assessing the classes of its K closest neighbours.

5. Comparison of Algorithm

The performance of three machine learning algorithms, namely Logistic Regression, K-Nearest Neighbors (KNN) with the IBK method, and Random Forest Classifier (RFC), was systematically compared across multiple dimensions. The evaluation criteria encompassed accuracy, speed, robustness, scalability, and interpretability. These analyses were conducted using a diverse set of student data samples, each with varying numbers of instances [3].

Accuracy and scalability

The comparative assessment of Logistic Regression, KNN (IBK), and Random Forest Classifier (RFC) was carried out with a keen focus on accuracy, processing speed, robustness, scalability, and interpretability. To ensure comprehensive analysis, three distinct sets of student data samples were employed, each exhibiting a unique distribution of instances.

Model	Accuracy
Logistic Regression	0.8990
KNN	0.9082
Random Forest classifier	0.9908

Figure 15: Comparison of the result between different ML models.

In evaluating the accuracy of different models from figure 14, we find notable distinctions. The Logistic Regression model showcases a commendable accuracy of 0.8990, indicating its effectiveness in making precise predictions. The K-Nearest Neighbors (KNN) model slightly surpasses this accuracy with a score of 0.9082, further solidifying its reliability in predicting outcomes. However, the Random Forest classifier stands out prominently, boasting an impressive accuracy of 0.9908, signifying an exceptionally high level of accuracy and precision in its predictions. The Random Forest classifier demonstrates superior predictive capabilities compared to Logistic Regression and KNN, making it a compelling choice for this dataset. These accuracy metrics provide valuable insights into the comparative performance of these models, aiding in informed decisions for selecting the most effective predictive model.

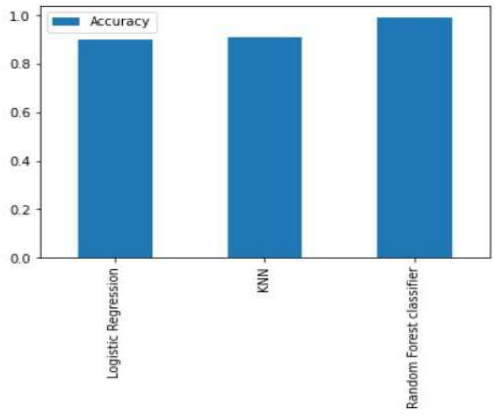

Figure 16: Comparison between the accuracy of Different ML models.

The results, as illustrated in the graph image above, clearly depict the performance of these algorithms in terms of accuracy. Notably, the Random Forest Classifier emerges as the front runner in terms of accuracy, showcasing its robustness and effectiveness in handling the dataset.

5.1 Speed and robustness

The comparison of model building and testing times between Logistic Regression and Random Forest Classifier (RFC) yields valuable insights into their computational efficiency. Logistic Regression exhibited remarkable efficiency by taking zero seconds for both model construction and testing on the split test dataset. In contrast, RFC required 0.94 seconds to build the model but performed the testing phase instantaneously. The time needed to train a model is commonly referred to as the build-time complexity.

This build-time complexity depends on the dataset's characteristics, typically denoted by 'n' points and 'd' dimensions. In the case of Logistic Regression, the built-time complexity is expressed as O(n*d), signifying a linear relationship between the number of data points and dimensions.

In contrast, the built-in complexity of RFC is represented as O(n*log(n)dm), where 'm' accounts for aggregation. This indicates that RFC's build-time complexity is significantly higher and more time-consuming, particularly as 'n' and 'd' increase in size.

Consequently, it becomes evident that RFC's primary computational cost lies in the model-building phase, making it the most time-intensive aspect of the model development process[5].

6. Conclusion

In this study, we leveraged a dataset containing student records from the Vivekanand Education Society's Institute of Technology. We divided the students into two distinct groups: one received traditional teaching methods, while the other experienced the POGIL teaching approach. We meticulously recorded the students' performance in various critical exams, including HSC, SSC, JEE, and MHTCET. Our comprehensive analysis of this dataset unveiled a noteworthy trend: students instructed using the POGIL approach consistently outperformed their counterparts who were exposed to traditional methods.

One significant observation was that POGIL not only contributed to improved academic outcomes but also fostered essential communication and teamwork skills. The collaborative nature of POGIL, where students work in small groups, encouraged them to articulate their ideas, pose questions, and collectively solve problems. This mirrors real-world scenarios where effective collaboration is indispensable.

Furthermore, we employed diverse algorithms to evaluate the data, specifically comparing Logistic Regression, KNN, and Random Forest Classifier. Among these, the Random Forest Classifier stood out with an impressive accuracy rate of 99.08%, making it the top-performing algorithm. Notably, this algorithm excelled among various tree-based algorithms.

As the dataset size increased, the J48 algorithm demonstrated superior performance and speed compared to other algorithms. In terms of robustness, all algorithms exhibited commendable performance. It's crucial to note that the choice of the most suitable algorithm depends on the specific application, considering factors like scalability, robustness, and the desired level of accuracy required for the application at hand.

In conclusion, this research highlights the efficacy of the POGIL teaching approach and the pivotal role it plays in enhancing both academic performance and critical interpersonal skills. Furthermore, the evaluation of machine learning algorithms underscores the significance of selecting the right tool for the task at hand, emphasizing the importance of scalability, robustness, and accuracy in various applications.

7. Recommendation

Our research strongly recommends that other researchers employ RFC, Logistic Regression, J48 decision trees, and the IBK algorithm for data mining purposes. These algorithms have proven highly effective in accurately measuring students' performance. By utilizing the right predictor attributes, which contribute significantly to the analysis of students' learning, we can derive insights that hold the potential to enhance students' academic achievements.

From this research, it becomes evident that a student's academic performance history is just one of several factors that influence their success at the university. Other critical determinants include parental education level, economic background, the home environment, occupation, physical disability, and health-related issues. While some of these factors may be noted on the admissions form, they are often not recorded in the database. As a result, we strongly recommend that the admissions team consider incorporating these additional details into the database. This proactive step will enable more comprehensive research in future studies, offering a holistic view of the factors impacting student success.

References

[1] Sandhya Kode EnhanceEdu Jyotsna CherukuriVNR Vignana Jyothi Institute of Engineering and Technology Creating a Learner Centric Environment Through POGIL Our experience in engineering and management education in India

[2] S. DE GALE, L. N. BOISSELLE Science Education International Vol. 26, Issue 1, 2015, 56-61 The Effect of POGIL on Academic Performance and Academic Confidence.

[3] Faiza Umar Bawah Department of Computer Science, Kwame Nkurmah University of Science and Technology, Najim Ussiph, PhD Department of Computer Science Kwame Nkurmah University of Science and Technology, Appraisal of the Classification Technique in Data Mining of Student Performance using J48 Decision Tree, K-Nearest Neighbor and Multilayer Perceptron Algorithms.

[4] Ms.A.Pavithra,andMr.S.Dhanaraj 2Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi Prediction Accuracy on Academic Performance of Students Using Different Data Mining Algorithms with Influencing Factors.

[5] Dogan, N. and Tanrikulu, Z. (2013). A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. Information Technology and Management, 14(2), pp.105-1

[6] Thirumuruganathan, S. (2010). A detailed introduction to the K-nearest neighbor (KNN) algorithm. Algorithm. https://saravananthirumuruganathan. WordPress. com/2010/ 05/17/a-detailed-introduction-to-k-nearestneighbor-knn-algorithm/ Accessed on 30th April 2016

[7] Baradwaj, B.K. and Pal, S. (2012). Mining educational data to analyze students' performance. International

[8] Journal of Advanced Computer Science and Applications. (IJACSA), Vol 2 No. 6 pp 63-69

[9] Image reference: Javatpoint.

[10] Soham Navandar, Ganesh Bhutkar, A Review on Data Exploration and Data Mining Evolution 2022 IJCRT | Volume 10, Issue 10 October 2022 | ISSN: 2320-2882

[11] Rohmah, Y. N., &Muchlis. (2013). Application of learning with POGIL strategy on soluble material and solubility times to train Kemampun critical thinking of students of Class XI SMA Negeri 1 Sooko Mojokerto. Une sa Journal of Chemical Education, 2(3), 19-23 [accessed May 25, 2018].

[12] Simonson, S. R., & Shadle, S. E. (2013). Implementing process oriented guided inquiry learning (POGIL) in undergraduate biomechanics: Lessons learned by a novice. Journal of STEM Education: Innovations and Research.

[13] Zawadzki, R. (2010). Is Process-Oriented Guided-Inquiry Learning (POGIL) suitable as a teaching method in Thailand's higher education? Asian Journal on Education and Learning, 1(2), 60-74.

[14] Zamista, A. A. (2016). Influence of learning model Process Oriented Guided Inquiry Learning to the skills of the process of science and student cognitive in the subject of physics. EDUSAINS (Online).