[[1]](#footnote-0)

**POGIL data analysis employing J48 Decision Tree, K Nearest Neighbor, Random Forest Tree, and Logistic Regression Algorithms to examine student performance**

First A. Author, *Sahil Parab*, Second B. Author, Priyanshu Singh, and Third C. Author, Mayur Pimpude Fourth Author. Anjali Yeole, Fifth E. Author, Maya Bhatt

***Abstract*—**Education is the acceptance that education improves the potential of each person. Each individual is unique and elevates the person academically, significantly, and passionately. It is significant for achieving life realization and for self-improvement**.** This study uses the classification techniques of data mining to mine data of students of Vivekananda Education Society Institute of Technology, Chembur, Mumbai to ascertain if there is any pattern between the grades where Students learned with POGIL and without POGIL. POGIL (Process Oriented Guided Inquiry Learning) is a cooperative learning method that integrates Guided Inquiry into a cyclical system of concept generation, investigation, and application. Final the test revealed an enhanced score among the students learned using the POGIL method

**General Terms**—Data analysis, classification

**Keywords**—J48 decision tree, K-nearest neighbour, Random Forest Tree, logistic Regression, WEKA

# Introduction

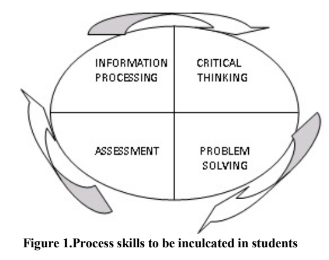
A

student-centered, group-learning instructional technique and philosophy known as Process Oriented Guided Inquiry Learning (POGIL) was created as a result of research on the most powerful strategies for learners to study. POGIL was Developed in 1994 to improve general chemistry

instruction. Today, more than 1,000 American high schools and institutions use POGIL in a variety of areas. The design of a POGIL exercise must consider these two factors above everything else. First, in order for students to generate the desired notions during the initial "Exploration” c presented. The guiding questions must also be properly prepared in

order for students to not only arrive at the correct conclusion but also to acquire a variety of process

and learning abilities. The first few questions usually expand on the knowledge that students already have and draw attention to the details that the model has to offer. Questions meant to encourage the identification of correlations and patterns in the data and eventually some concept development follow.  The last few questions can 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. Constructivism serves as a solid foundation for POGIL instruction.[1]



**Figure 1 :Process skills to be included in students.**

The purpose of analysing student performance is to Ascertain if there is any pattern between values and relationship among data entities where Students learned with POGIL. Many algorithms, including Naive Bayes, J48 decision trees, K-nearest neighbour, and others, are used in data mining to analyse student performance and extract knowledge, such as classification. This information can help forecast student progress by revealing significant correlations and unexpected outcomes. [1]

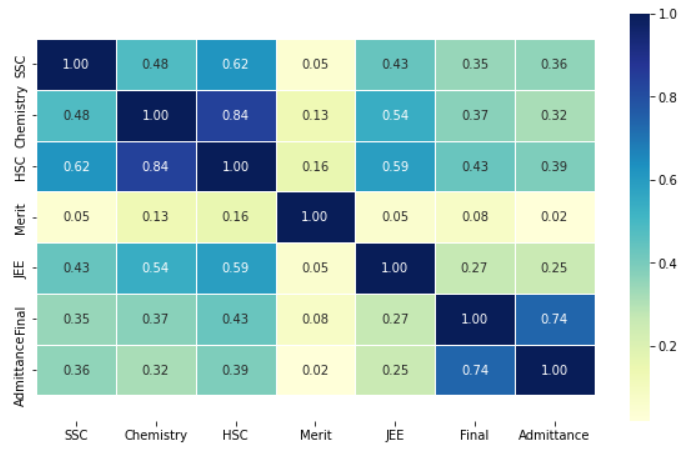
# About Student dataset

We have two dataset – data with POGIL and data without POGIL.

In which dataset with POGIL have POGIL attribute along with JEE, SSC, HSC, Admittance, Final and Merit attributes which consist of 118 values of students details.

The dataset consists of 545 student’s details and there are 7 column’s those are JEE, SSC, HSC, Admittance, Final and Merit. On working with this dataset, it has been known that their parents are mostly working as an agriculturist, belongs to annual income of 50,000 and below.

The students need to concentrate little better in their studies according to the observation.



**Figure 2: Confusion Matrix**

Through this correlation heat map, we can find out the most inter related attributes from the dataset provided. As we can observe there is good relation between final and admittance attribute also relation can be mapped between HSC-chemistry, SC-chemistry, HSC-SSC.

**What is the Effect of POGIL on the Academic**

**Confidence of Students?**



**Figure 3: Summary of Data Set**.

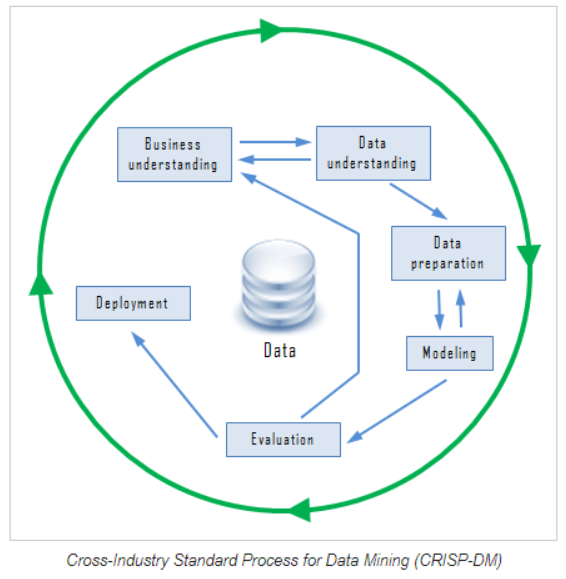
# Data Mining

Finding patterns and other important information from huge data sets is a technique known as data mining, commonly referred to as knowledge discovery in data (KDD). Data mining is described in one of the most comprehensive ways by Gartner Inc., who defines it as "the process of discovering new, meaningful correlations, patterns, and trends by sifting through enormous amounts of data stored in repositories and by utilizing pattern recognition technologies as well as statistical and mathematical techniques."

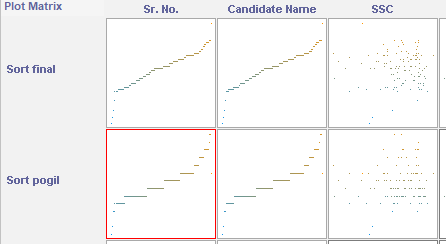
With smart data analytics, data mining has improved corporate decision-making. These analyses' underlying data mining techniques can be classified into two categories: those that describe the target dataset or those that forecast results using machine learning algorithms. The most interesting information, including fraud detection, user habits, bottlenecks, and even security breaches, are surfaced using these approaches for organizing and filtering data. In an emerging field called "educational data mining," methods are being developed for examining the distinctive and constantly expanding amounts of data that are generated in educational settings, with the goal of using these methods to learn more about students and the environments in which they are taught.

# 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). 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. Modelling include 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.

**

**Figure 4: Data mining chart**.



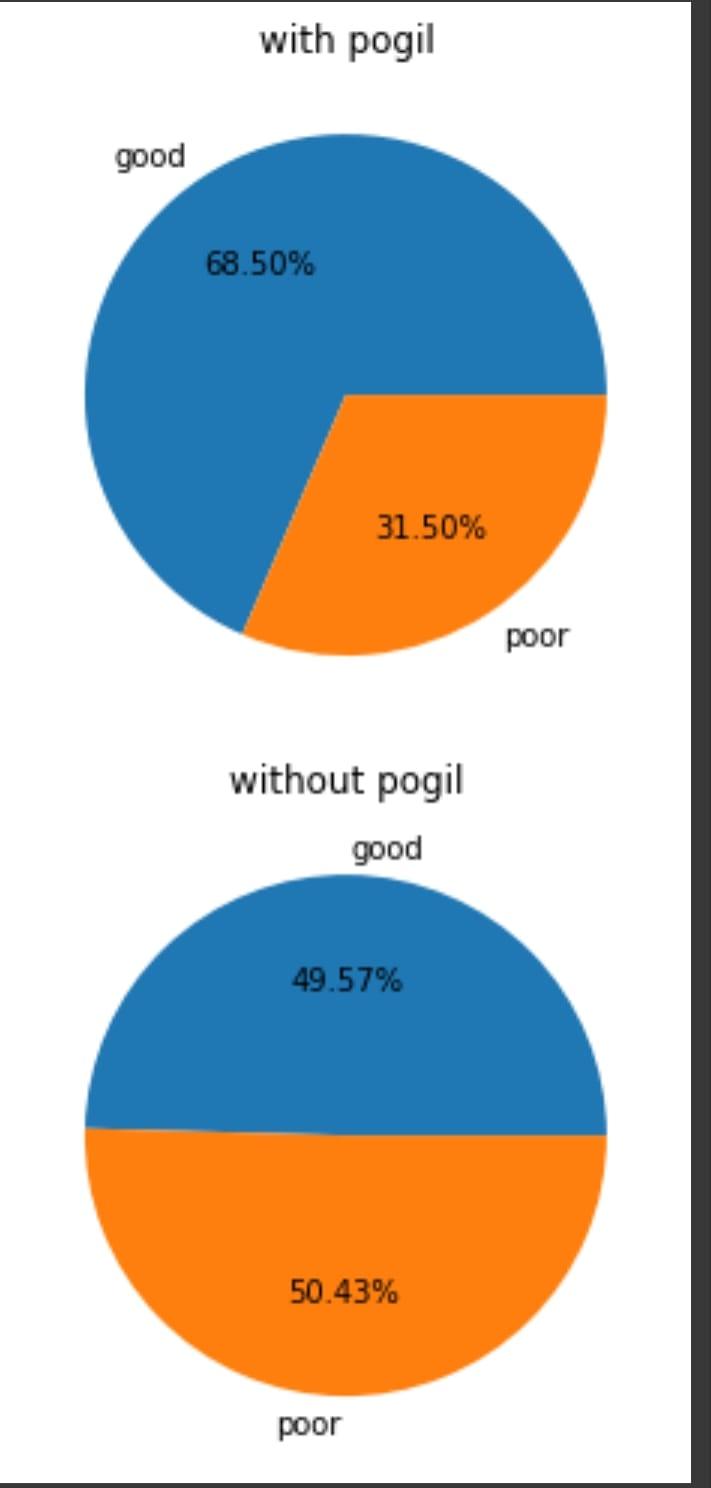
**Figure 5 : Data Analysis .**

*Waikato Environment for Knowledge Analysis (WEKA)*

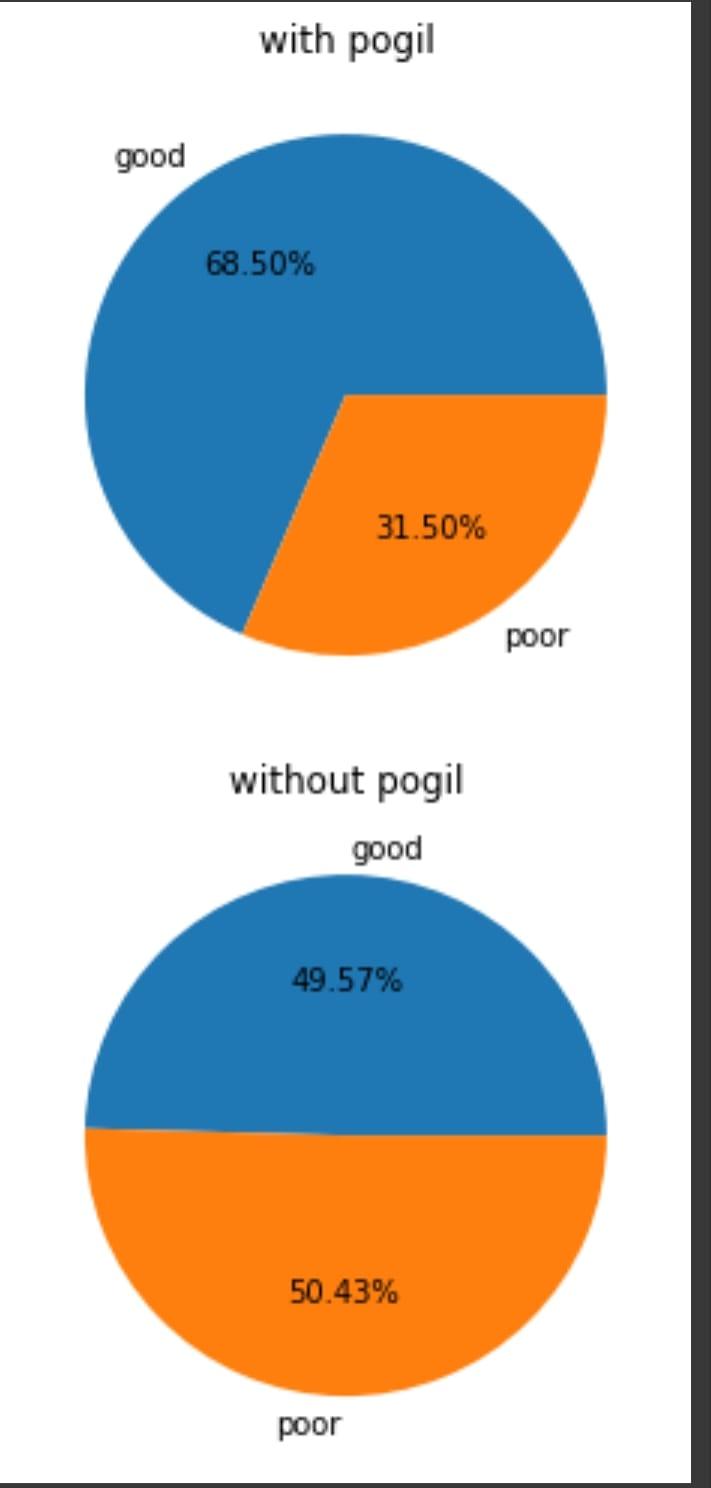
The University of Waikato in New Zealand developed the WEKA software written in Java which was used for this research. WEKA supports a wide range of algorithms and large data sets. It is an open-source software issued by GNU General Public License. It contains tools for data-pre-processing, algorithms for classification, clustering, regression, visualization, and association rules. Basically, used for data analysis and predictive modelling. It has a graphical user interface making it easy to access the various functions.[3]

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

To use we have uploaded out csv dataset to WEKA for analysis.



**Figure 6: Results after using Pogil technique.**



**Figure 7 :Result without using pogil technique .**

#### classification

Classification is a technique where we categorize data into a given number of classes. It is a supervised learning method that requires a labeled dataset to train and build the model. The classification model classifies to which the class the input values belong to which the model is trained on data.

It will predict the class categories for the new data. A feature is an individual measurable property of a phenomenon being observed.

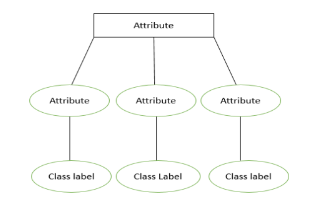


Classification Algorithms could be broadly classified as Linear classifier, support vector machines, Quadratic classifier, kernel estimation, Decision trees, Neural Networks, Learning vector quantization

The approach is based on machine learning. The classification technique classifies items in a data set into a pre-defined set of groups. Data classification entails learning and classification algorithm and in classification, test data is used to estimate the accuracy of classification rules if the accuracy is accepted the rules are then applied to the new data set. [3]

#### *C4*.5 algorithm/J48

The C4.5 algorithm is a classification algorithm that produces decision trees based on information theory. It is an extension of Ross Quinlan’s earlier ID3 algorithm also known in Weka as J48, J standing for Java. The decision trees generated by C4.5 are used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

The J48 implementation of the C4.5 algorithm has many additional features including accounting for missing values, decision tree pruning, continuous attribute value ranges, derivation of rules, etc. In the Weka data mining tool, J48 is an open-source Java implementation of the C4.5 algorithm. J48 allows classification via either decision trees or rules generated from them.

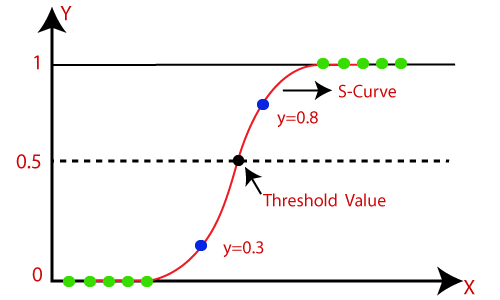
**Figure 8: Decision Tree.**

This algorithm builds decision trees based on a set of training data in the same way the ID3 algorithm does, by using the concept of information entropy. The training data is a set S= {s1, s2, …} of already classified samples. Each sample si consists of a p-dimensional vector (x1, i, x2, i, …, xp, i) where the xj represents the attribute values or features of the corresponding sample, as well as the class in which the sample in which the sample falls. To gain the highest classification accuracy, the best attribute to split on is the attribute with the greatest information.[3]

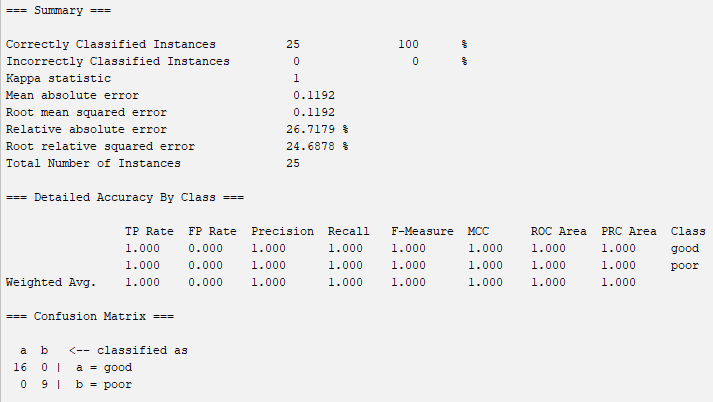
At each node of the tree, the C4.5 algorithm chooses the attribute of the data that most effectively splits its set of samples into subsets, enriched in one class or the other. The splitting criterion is the normalized information gain, which is calculated from the difference in entropy. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the partitioned sub lists utilizing a divide-and-conquer approach and creates a decision tree based on the greedy algorithm

#### 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. Logistic regression uses the function called the sigmoid function. 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 9: Logistic Function.**

****

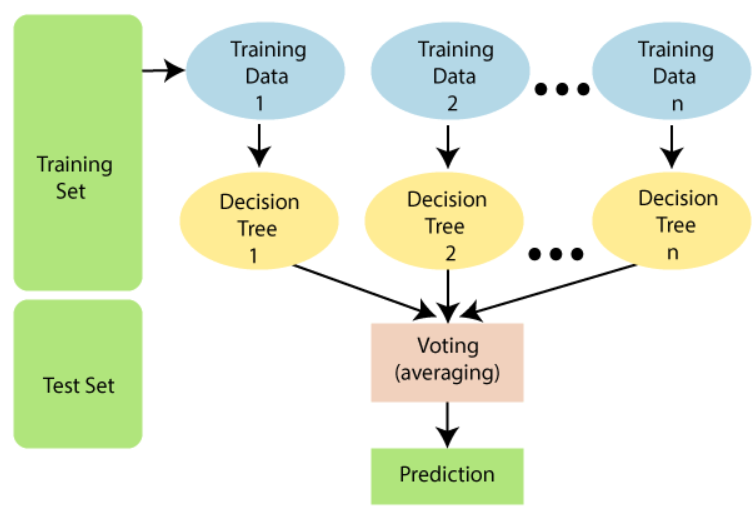
**Figure 10: Result of Logistic Regression .**

#### Random Forest Tree

Random forest classifier is a supervised machine learning technique. It can be used for both classification and regression problems in machine learning. 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.

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 final 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.

.

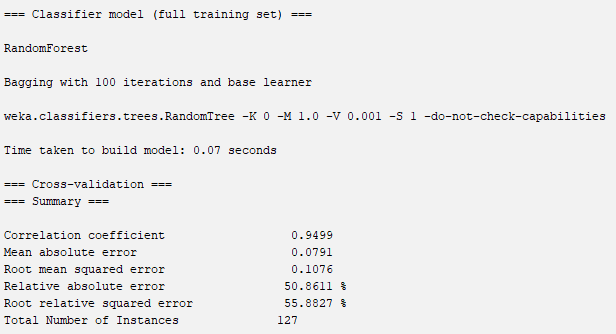


**Figure 11: Random Forest Tree.**

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, there are two assumptions for a better Random Forest classifier:

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

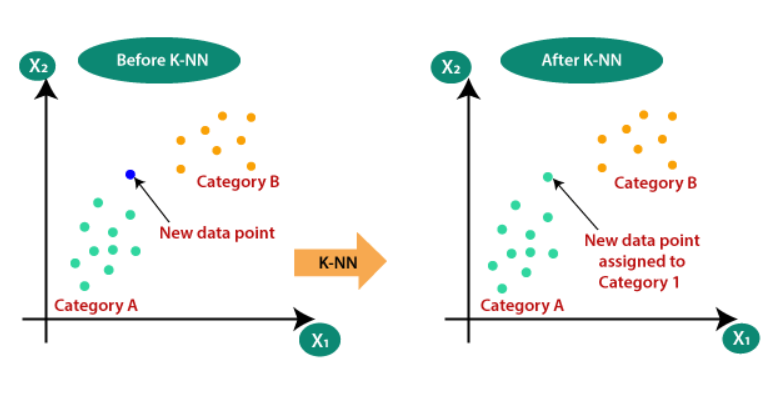
The predictions from each tree must have very low correlations.



**Figure 12 : Result analysis of Random Forest** .

#### K Nearest Neighbor

K Nearest Neighbor is a simple machine-learning algorithm of supervised learning. It can be used for regression and classification problems of machine learning. 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 it checks the similarity between them and put them in a special category. It is also called the lazy learner algorithm because it does not learn at the time of training it classifies the dataset at the of prediction. 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 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.



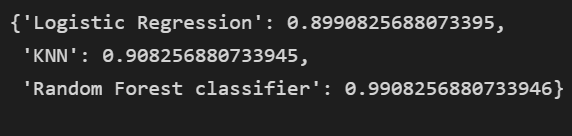
**Figure 13: KNN Clustering.**

# Comparison of algorithm

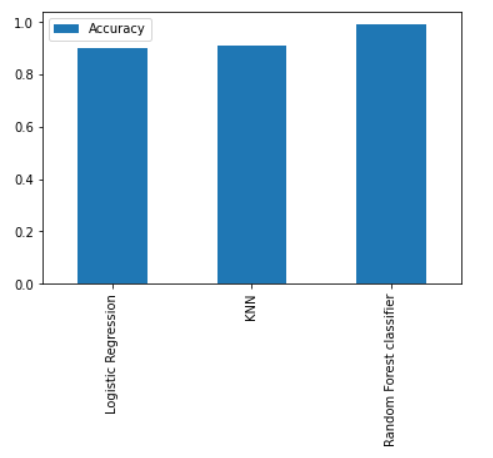
The performance of Logistics Regression, KNN (IBK) and Random Forest Classifier (RFC) were compared based on accuracy, speed, robustness, scalability and interpretability of the algorithms. Three student data samples were used. Each data had a different number of instances.

### Accuracy and scalability

RFC performs significantly better than IBK and on all data sets. The classifier accuracy increases with lager data set. While for IBK and RFC, the classifier accuracy decreased for student sample 2 and increased for student sample 3. Logistics Regression had a classifier accuracy of 89.90%, IBK had a classifier accuracy of 90.8256% and RFC had a classifier accuracy of 99.0825%.



**Figure 14 : Comparison of the result between different ML model.**



**Figure 15: Comparison between the accuracy of Different ML model.**

From above graph image we can see the accuracy of three algorithms in which random forest classifier has highest accuracy.

### Speed and robustness

Logistic Regression and IBK took zero seconds to build model and test the model on test split while RCF took 0.94 seconds to build model and zero seconds to test model on test split.

Built time complexity is the amount of time taken to train the model.

Assuming dataset has ‘n’ points and ‘d’ dimensions.

* Built time complexity of Logistic Regression is O(n\*d).
* Built time complexity of IBK has no built time.
* Built time complexity of RFC is O(n\*log(n)\*d\*m) (where ‘m’ is aggregation).

From this we can say that RFC has highest time complexity of building a model.

**CONCLUSION**

There is no single answer as to the best algorithm that should be used to mine student data. The algorithm that would be used highly depends on the task to be performed. J48 algorithm produced the best classifier and predictor models with 99.08% accuracy but it could not predict the exact values.

It had the greatest difference between predicted and actual values. IBK had the least difference between its predicted values and the actual. While MLP had a relatively poorer predictor accuracy but with better difference between actual and predicted values.

J48 decision tree performs better with increasing data size as the classifier accuracy increases followed by IBK and then RFC. J48 and IBK are faster at building and testing the model as compared to RFC. In terms of robustness, all algorithms handled the noise better since the percentage of accuracy did not reduce drastically.

J48 had the best interpretability as it correctly classified all instances. Amongst the three algorithms J48 decision tree is the best even though the other algorithms are good enough for analyzing and predicting student performance.

**RECOMMENDATION**

We strongly recommend that other researchers adopt RFC, Logistic Regression, J48 decision tree, and IBK algorithms to mine students’ performance using the WEKA tool to explore more predictor variables that contribute to the choice of parents/guardians, students, and higher education institutions to improve quality of education and national development.

Based on this research, it is clear that academic performance history is only one factor of a student amongst many that determine the success of a student in the university. Other factors that help determine the performance of students are parents’ level of education, economic background, the environment at home, occupation, physical disability, and health-related issues. These are usually captured in the admission form but not entered into the database. It is therefore recommended that the admission team input such details into the database for more holistic research to be conducted in future works.

# References

1. Sandhya Kode EnhanceEdu Jyotsna Cherukuri

VNR Vignana Jyothi Institute of Engineering and

Technology Creating a Learner Centric Environment Through POGIL Our experience in engineering and management education in India

1. 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.
2. 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.

1. Ms.A.Pavithra ,and Mr.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.
2. 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
3. Thirumuruganathan, S. (2010). A detailed introduction to 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
4. Baradwaj, B.K. and Pal, S. (2012). Mining educational data to analyze students' performance. International
5. Journal of Advanced Computer Science and Applications. (IJACSA), Vol 2 No. 6 pp 63-69
6. Image reference : Javatpoint .

1. [↑](#footnote-ref-0)