A Data Mining Technique: Predicting Students Performance, and Future

Hadisur Rahman[1], Md. Meraj Ali[2], Md. Mainuddin[3], Md. Sumon Rony[4], Md. Tanvir Hossain Badhan[5]
Department of Computer Science and Engineering, Varendra University[1, 2, 4, 5]
Florida State University[3]
hudacse6@gmail.com, meraj09034@gmail.com, mainuddincse@gmail.com, sumoncse2395@gmail.com
imtanvir.cse@gmail.com

***Abstract* — Nowadays, it's so hard to anticipate a student future after finishing their academic study. Sometimes, they even don't know what they are going to do when they finished their study. Now, in the competitive world, it's not so easy to get a good job. Today's highly challenging educational field there is no more scope to avoiding this system. This because of our variation of the intricate educational system. This system makes our job section more challenging. When we are trying to get a good job, the job seeker first checks from which educational institution we came from. Now, it's a common tend to our job market. Mostly, even, they emphasized educational institution from where we are belonging rather than the experiences we have. Then, they checked the other necessities they needed to be. Sometimes, many of the student's desperately do suicide about just thinking their floating future. This is the main reason for this research, this research tells us about the future of a student and also the lacks, the failures, and the maladroit so that we took necessary steps for what should we do need to make their future much better than they have before and how to come up with the problems they have. And we also get to know about the fissures of our educational system.**

Keywords— Data Mining, Algorithm, Classifier, Prediction, Methodology, Job, Dataset, & unsupervised and unsupervised algorithm.

I. INTRODUCTION

Data mining [2, 6] is a mathematical method of processing data which is strongly tested in varies areas that intent to get useful information from the data. Here we collected 17 type of data based on a student's whole profile (for both who get a job already and who are trying to find a job) like CGPA, is we done thesis/project in the last year, Programming skills, General Knowledge Skills, Class Attendance, how much is we active web/programming/robotic club, like to read book/playing/journey nor not, like to work any environment [3] or not and etc. Then we follow the CRISP/DM2 in our work through and we used or applying several data mining process, algorithms, classifier as like (k-NN, ZeroR, KStar, Multilayer Perceptron, Naive Bayes, j48 and etc.) in pruned or unpruned dataset along with varies instances. Then we visualize the data to check the percentage of error rate, correctness and to made decision in each and every step and process. And we also pretty forward to see which of the following data mining algorithm or classifier working better and also trying to apply by our own code and checking is it work properly or not.

After process and visualizing our data we classified the result by dividing some field as like: is the student needs to be more academic study or finishing study but still he/she needs more improve, which section did he/she needs to improvement and what kind of skills does need mostly to get a proper job after finished study. In our total collection of instances as an example: 106, After analyzing this data we came to know that 33 students Get their job after finishing their academic education, 18 students do not get a job, 31 Student's need more improvement and 24 students still need to academic study.

So, by analyzing data result we try to develop our system more efficient way to get more job in our instance's. To get this aggregate result we apply several data mining approaches using data mining algorithms, classifier each and every algorithm's and classifier produces a different result but we have shown here the most average correct errorless aggregate result. In this research, a student has been classified about their success in the job sector as they what to do need by using their previous activities on the academic educational environment to get an appreciable job with highly lives. It also can be summarizing that this approach can be used to help students and to improve their performance; by reduce or failing ratio by taking rightfully steps at exact time to progress the quality of learning. As learning is an active method, interactivity is a rudimentary element in this process that affects students' satisfaction and performance [4, 5]. For future work, the experiment can be extended with more broad distinctive attributes to get more accurate results, used to improve the students learning outcomes. Also, experiments could be done using additional data mining algorithms to get an immense approach, and more valuable, and accurate outputs. Some varies software may be utilized whereas the same time various ingredient will be used.

II. METHODOLOGY AND EXPERIMENT

A. *Analysis Tools*

Waikato Environment for Knowledge Analysis (WEKA) [1] is a notable suite of machine schooling which is penned in Java and developed at the University of Waikato. It is free software [Fig. 1.] approachable under the General Public License (GNU). It contains alms of algorithms and visualization tools for forecast modeling, data analysis, along with graphical user interfaces for flush access to this functionality. It supports various standard data mining tasks, more conspicuously, data pre-processing, classification, visualization, clustering, feature selection, and regression.

Fig. 1. Weka Software

B. *Data Collection and format*

Data is collected from observing of 106 students in Dept. of CSE of Varendra University, Rajshahi, Bangladesh in April

2017. We categorized the data by 17 types for predicting knowledge (Student future) to get the result based on this data. The 17 types of data are lookalike:_CGPA, done thesis or project, class attendance in total 12 semesters, class assignment in total 12 semesters, class presentation in total 12 semesters, healthy, Do-programming, was he/she active in programming club, was he/she active in web development club, General knowledge club, Attitude, was he/she like reading Books, was he/she like to play, was he/she like to journey, need to retake, and is he/she like to work any environment [8, 17].

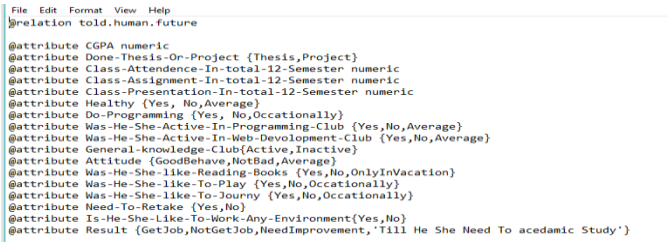


Fig. 2. Data Collection type and Format

C. *Workflow Processes*

It is very crucial to follow a correct process from the starting

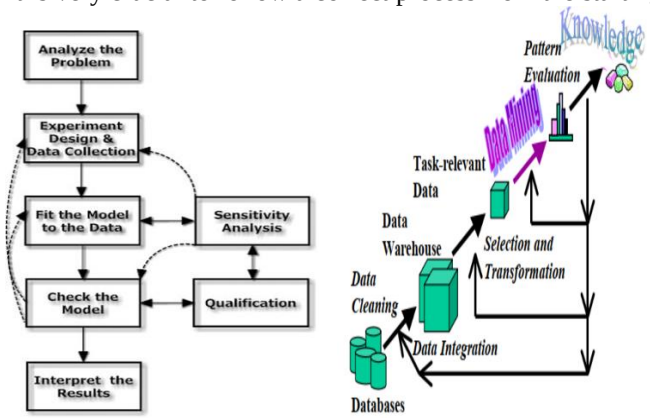


Fig. 3. Data Workflow & Knowledge Discovering process.

point until the creation and the deployment of the solution. We will use during this thesis project, the CRISP/DM2 (Cross Industry Standard Process for Data Mining) describes on the following figure: Problem understanding, data understanding, data Pre-Processing, modeling, evaluation, and deployment [14, 15, 12].

D. *Load Data into Weka*

As Weka only hold up the two types of format first one: ARFF (Attribute-Relation File Format), and second one: CSV (Comma Separated Values). As our input data is in jpg format so we need to transform it into CSV or ARFF. By transforming the jpg image to CSV format then we loaded it into the Weka.

E. *Data Pre-processing*

- Data preprocessing [9] is a data mining procedure that relates to data transforming in raw data into an accessible structure. Real-world data is almost certainly inaccessible, not coherent, and/or lacking in certain conduct or trends, and is probable to contain many err. Data preprocessing is a demonstrated method of rectifying such issues.
- We use six types of steps to pre-process our data: Import the libraries, import the data-set, check out the missing values, see the classification values, dividing the data-set into training and test set, and then feature scaling.

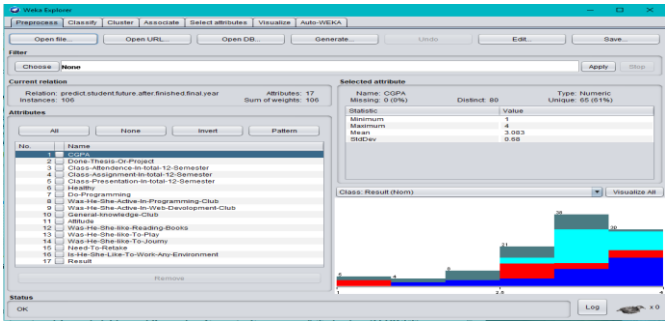


Fig. 4. Data Preprocessing into Weka

F. *Data mining method*

In plausibly data mining to be a mathematical procedure for pre-processing data which is successfully applied in various fields that, point to gain handy grasp from those data. The procedures which is applied in Data mining are used to create a structure following what is not known as data will try to identify the new data. Instead of origin, all data mining techniques view one common attribute: automated discovery of new connections and dependencies of attributes in the notice data [10]. To categorization the data into a class, we need to follow some steps. The algorithm is two types of groups:

[1] unsupervised algorithms and
[2] supervised algorithms.

If the mining is "unsupervised" or "undirected", the result is not particularly viewed in the data set: the task of unsupervised algorithm is to discover naturally inherent patterns in the data without the previous information about which class the data could belong, and it may not load any supervision (Cios, Pedrycz, Swiniarski and Kurgan, 2007). Oppositely, in unsupervised learning, no target variable to be learned is to recognize as such. Instead, the unsupervised learning algorithm finds for patterns and models among all the variables. The aim is to create a model not to reveal the data patterns to the set of input fields. Sometimes, the model manufacture by an unsupervised schooling algorithm used for forecast tasks even though it was not designed for such a job. A technique of clustering and association laws belongs to this group.

Supervised algorithms are those which use data within the forward not unfamiliar class to which data belong for building structure, and then on the basis of the constructed model forecast, the class to which is not known data will belong from. A method of categorization belongs to this group. Methods of data categorization perform a procedure of learning a function that visualizes the data into one of several predefined cases. For every categorization algorithm that is based on inductive learning, the input data set is given that consists of an angle of ascribing values and their corresponding class. The goal of a categorization procedure is to build a model which makes it straightforward to rectify future data points based on a set of specific characteristics in a robotic way. Such systems take a collection of cases as input, each belonging to one of a limited number of classes and described by its values as a fixed set of ascribes. As a result,

they take a classifier that can precisely predict the class from where a new case belongs. The most common way of categorizations is decision trees, induction rules or classification rules, probabilistic or Bayesian networks, neural networks, and hybrid procedures.
There are so many dissimilar classifiers in the articles and one cannot choose for the best one, because they differ commonly in many aspects such as learning output, amount of data for training, classification speed, robustness, etc. In this study, we scrutinize the impact of three algorithms for intelligent data analysis: C4.5, Multilayer Perceptron, and Naive Bayes. Classification structure is created by using these algorithms whose forecasting aim is to forecast the class (student's success) to which some new not level sample will belong. The selected three categorization techniques are used to reveal the most suited way to forecast student's success.

G. *Applying Classifier's and Visualize this Classifier's*

In this stage we are applied all of precise data mining classifier's like k-NN, ZeroR, KStar, Multilayer Perceptron, Naive Bayes [11], and J48 [16]. And shown classifier's tree visualizations in [Fig. 6.] how it looks like when we apply all our collected data set on any kind of Algorithm or classifier. In [Fig. 5.] we have shown the obtain decision tree model for Naïve Bayes classifiers.

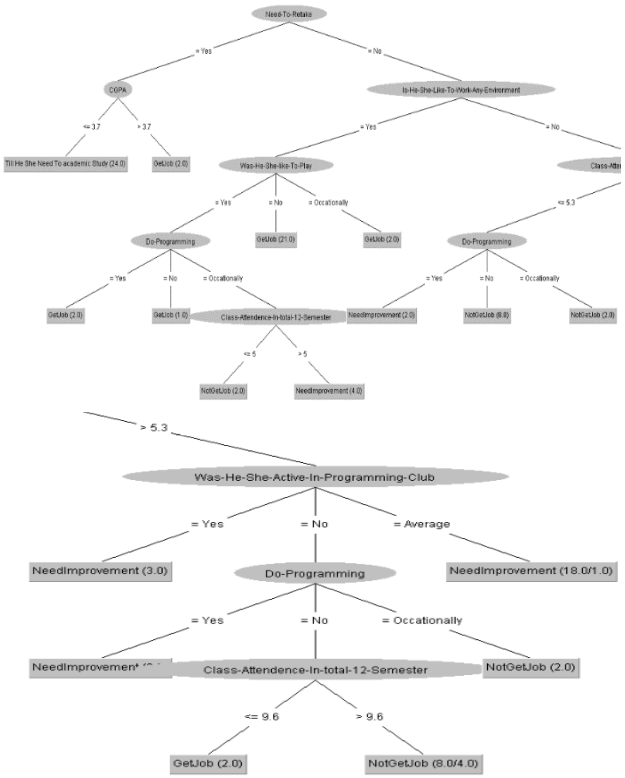


Fig. 5. Obtained decision tree model

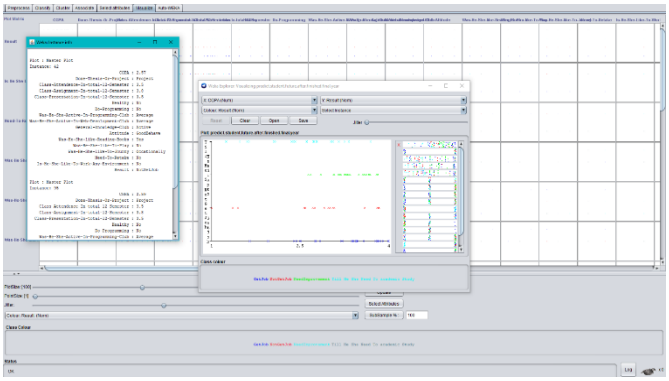


Fig. 6. Classifier's tree visualizations (Error prediction)

III. TESTING RESULT

The result had been classified by analyzing the data set. In conclusion, we come-up with that, 33(blue-color) students get their job after finished their academic education, 18(red-color] students did not get the job, 31(Aqua-color) student's need more improvement and 24(green-color) the student still needs to academic study [18].

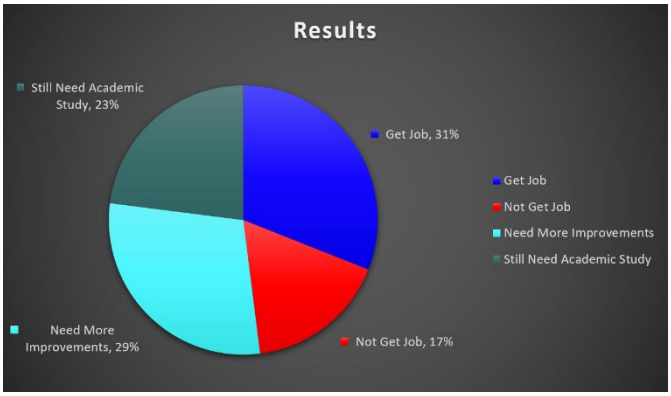


Fig. 7. Results

For generalize the results from those classifier's we used training and also unpruned false data set.

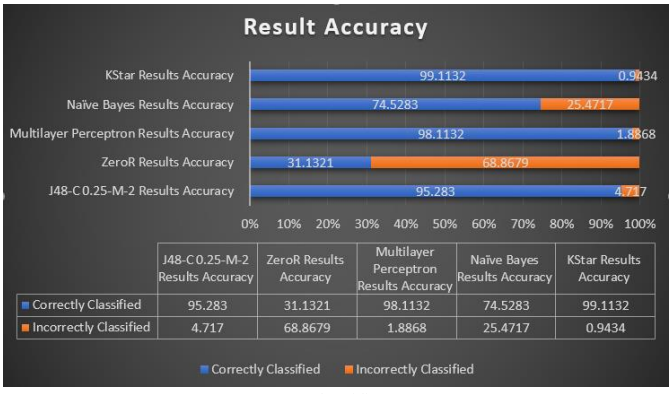
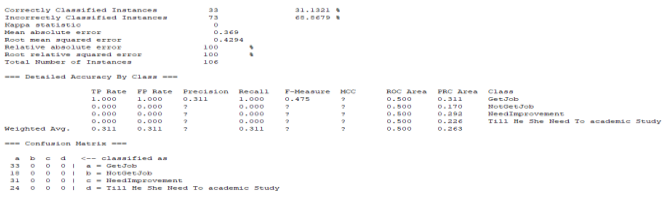


Fig. 8. Results accuracy on various classifier's

Here after we also provide some classifier's results that how much accuracy we got in from which classifiers.



```
Correctly Classified Instances          33               31.1321 %
Incorrectly Classified Instances        73               68.8679 %
Kappa statistic                          0
Mean absolute error                      0.369
Root mean squared error                  0.4294
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              106

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.311      1.000    0.475      ?        0.500     0.311     Get2Job
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.170     NotGet2Job
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.292     NeedImprovement
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.226     Till He She Need To academic Study
Weighted Avg.    0.311    0.311    ?          0.311    ?          ?        0.500     0.263

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 33  0  0  0 |  a = Get2Job
 18  0  0  0 |  b = NotGet2Job
 31  0  0  0 |  c = NeedImprovement
 24  0  0  0 |  d = Till He She Need To academic Study
```

Fig. 9. Results from ZeroR classifier's

VII. DISCUSSION & CONCLUSION

From our small data, the output of performance by selected classification algorithm are summarized and presented above. The achieved results reveal that the decision tree classifier (KStar) performs best--with the highest overall accuracy. The ZeroR classifiers are less accurate than the others. However, all tested classifiers are performing with an overall correctness above 80% which means that the error rate is low and the predictions are very reliable. To the extent that the comprehensive perfection for the varies classes is agitated, it is perceptible that the forecast is the best for some the Excellent class, and quite inferior for at the minimum one cases. The predictions for the Good and Very Good classes are more precise than for the other classes, and all classifiers perform with accuracies around 50-95 %. The decision tree classifier (KStar) is most reliable because they perform with the highest accuracy for all classes, except for the Excellent class. The ZeroR classifier is not able to forecast the classes which are less represented in the dataset.

After all pros and cons we can conclude that the data mining algorithms we were applied on the preoperative assessment data to predict the success of a student. The accomplishment of the learning procedure [13] were check out based on their forecast correctness, also the comfort of learning and user-friendly aspects. This schooling was based on conventional classroom environments since the data mining procedure was applied after the data was collected. However, this methodology might use to help students, and educator to enhance student's performance; reduce failing rate by taking apt steps the right time to enhance the aspect of the student future. As though something gaining is an active process, interactivity is the basic elements in this process that affects students' performance [7], and evaluation.

VIII. FUTURE WORK

For the future task, the experiment can be extended with more distinctive attributes based on the student future to get more accurate results to improve the students learning outcomes to make their future better [12]. Also, the analysis could be done using additional data mining algorithms getting an immense approach and more valuable and accurate outputs. Some different software may be utilized while same time various factors will be used. Also, it is important to answers these questions as a future task:

- How to obtain that predicting models are user-friendly for professors or non-expert users?
- How a student's class attendance, healthy or not healthy, and not attending a dept. club etc. influence to forecast the future?
- How to integrate data collection system of students and data mining tool?
- For a low-quality student what should be the penultimate to achieve the desired goal?
- How much predicting accuracy tend to fall in the student's actual and real life? How much is it correct with our providing future data model which has been predicted by us?

VIII. REFERENCES

[1] WEKA, https://www.cs.waikato.ac.nz/ml/weka/

[2] Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings, Cordoba, Spain.

[3] Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4) (2006) 255{268}.

[4] Wang, L., Hu, W., Tan, T.: Recent developments in human motion analysis. Pattern Recognition 36(3) (2003) 585{601}.

[5] Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3) (2001) 257{267}.

[6] Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[7] Superby J. F., Vandamme J.P., Meskens N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In International conference on intelligent tutoring systems, Educational Data Mining Workshop, Taiwan, 2006:1 – 8.

[8] Rohit Arora and Suman, Comparative Analysis of Classification Algorithms on Different Datasets using WEKA, International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012.

[9] Payal P. Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, "Preprocessing and Classification in WEKA Using Different Classifiers," Int. Journal of Engineering Research and Applications, Volume 4, Issue 8, August 2014.

[10] Sonam Narwal and Mr. Kamaldeep Mintwal, "Comparison the Various Clustering and Classification Algorithms of WEKA Tools, "International Journal of Advanced Research in Computer Science and Software Engineering IJARCSSE, Volume 3, Issue 12, December 2013.

[11] Tina R. Patil, Mrs. S.S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," International Journal of Computer Science and Applications, Volume 6, No.2, Apr 2013.

[12] Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

[13] Michie, D., Spiegelhalter, D.J. & Taylor, C.C. 1994. Machine Learning, Neural Statistical Classification, Ellis Horwood.

[14] W U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[15] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.).

[16] Rohit Arora, Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," International Journal of Computer Applications, Volume 54, Sept. 2012.

[17] Izzat AlsmadiandIkdam Alhami, "Clustering and classification of email contents," Journal of King Saud University Computer and Information Sciences, SCIENCEDIRECT, January 2015.

[18] Jaikee Kumar Singh, Shruti Pareek, "Comparison of Different Classification Techniques Using WEKA for Hepatitis," Indian Journal of applied research, Volume: 4, Issue: 10, October 2014.