Naïve Bayes Classifier for Predicting Success Factors of Engineering Education: Belgian Example

Saeka Rahman
Department of Computer Science and Engineering
International Standard University
Dhaka, Bangladesh
saeka@isu.ac.bd

Abstract

This research attempts to understand the underlying factors influencing the success of engineering education. In doing so, it analyzes the available data regarding the non-mandatory positioning test, which was introduced at the Belgian universities for engineering programs in summer 2013. Predictive model learning algorithms are used to make prediction for unseen data. In this research Naïve Bayes based algorithm is used to predict the contributing factors for the success in engineering education. The result shows that the prior academic achievement (choosing option of higher hours in mathematics, percentage of marks in mathematics in high school) of the students influences the score of the test. It also shows that the score of the test along with prior mathematical experience is a good predictor for the success and failure of students in engineering education. Moreover, the research finds that the test score has a high predictive power for the result of engineering study especially for the students who are more likely to do badly. On the other hand, the results indicate that gender is not an obstacle in study success in engineering education. This study finds that the girls who do conduct higher hours of mathematics in their high school and go for the engineering study do equally well as boys.

Keywords

Engineering education, Women in engineering, Predictive model learning, Naïve Bayes classifier.

1. Introduction

The technological advancement in the wave of Fourth Industrial Revolution is creating opportunities of a wide range of entirely new jobs and reducing the number of workers required for certain work tasks (CNES 2018). Therefore, engineering practice, research, and education are of great importance to cope with future challenges regarding economy, environment, energy, health, and security (Duderstadt 2010). In EU countries engineering employment is projected to grow by 15 per cent over the period from 2018 to 2030. Around 4.3 million job openings will need to be filled in this time period. This trend is expected to be sustained in the following years (Skills Panorama 2019 update).

However, the number of engineers needed is not enough to meet the demand. Still there is deficit in the supply side in Science, Technology, Engineering and Mathematics (STEM) workforce (Skills Panorama 2016). Low retention rate in this discipline of study is an international concern for a long time (Steenkamp et al. 2017). Moreover, though this education field is vital in national economies, the gender gap between men and women in this field is globally apparent (UNESCO 2017). High dropout in engineering study, less interest in science courses in high school, and negative perception about STEM education and career especially among the women are considered as shortage of workforce in this field (EU Skills Panorama Analytical Highlight 2012, Hill et al. 2010). Therefore, in order to improve the retention and graduate rate and to encourage more female participations, it is important to determine the contributing factors that influence the academic success in engineering education (Hall et al. 2011).

In order to predict the success factor of engineering education, the research analyses the data regarding positioning test for engineering education in Belgian Universities. As presented in Vanderoost et al. (2014), for more than hundred years admission to engineering studies at the universities of Belgium was based on an entrance exam. This multi-topic exam was used to judge the high school mathematical skill required for engineering problem solving. However, despite academic opposition, this exam was abolished by the government in 2004. Onwards 2004, anyone who has a high school degree is eligible to study engineering without any subject specific requirements. A basic course on mathematics is introduced in the first semester in order to maintain the required level of math skill. Researches show

that the success rate decreased by more than 21% compared to the time of the entrance exam (Vanderoost et al. 2014). This creates problem for teachers and students due to the difference between expected and actual prior skill (ATTRACT Project 2012, p.129). The mathematical skills in general are considered as a high predictor for success in engineering. Therefore, to judge the prior mathematical skill a non-mandatory positioning test is introduced at the Belgian universities in summer 2013 (Vanderoost et al. 2014).

1.1 Objective

The objective of the research is to analysis the available data regarding the positioning test in order to discover important features explaining the result of the positioning test (hours of mathematics in prior education, gender, percentage in mathematics) and its impact on bachelor study progress. In so doing Naïve Bayes algorithm is employed to measure the predictive ability of the test.

2. Literature Review

The goal of the predictive learning is to make prediction for the unseen data. The most used method is the attribute-value learning method. In this method a model is learnt to predict the value of one specific attribute (target attribute) from the value of other attributes. When the task is to predict the target attribute of unseen examples from a set of possible target attributes then it is called classification problem (Mitchel 1997, Blockeel 2013).

The most important part before constructing a classifier is the attribute selection. Attribute selection means the selection of relevant attributes and removal of redundant and/or irrelevant attributes. This technique helps to construct simpler and faster models. Attribute selection also provides the structural knowledge by knowing which attributes are inherently important to the application. Information gain attribute ranking is one of the simplest and fastest attribute selection method which is presented in Hall and Holmes (2003). This method provides an additional information gain about the class, given a particular attribute (Hall and Holmes 2003).

Moreover, it is very important to guarantee the generality of a classifier. The classifier will provide good results not only for observed data but also for unobserved data. In order to ensure generality of the classifiers, k-fold crossover validation technique is used. The data is repeatedly split into k folds in this technique. Among them (k-1) fold is used for training sets (set of labeled data) and one-fold is used as validation set (set of unseen data). The network is trained using the training sets and the resulting network is evaluated by the validation set. The average performance of the network over the validation set provides the quality. Typically, 10-fold crossover validation is used (Bouckaert 2008).

A Naive Bayes classifier is a simple and popular classification method based on Bayes rule and assumption of conditional independence of the attributes given the class (Bouckaert 2008 and Blockeel 2013).

Given the attribute values the most probable value of the class is:

$$v^* = argmax_{v_{j\in V}}\, p(v_j|a_1, \dots, a_n)$$
$$= argmax_{v_{j\in V}} p(a_1, \dots, a_{n,}|v_j)p(v_j)/p(a_1, \dots, a_n)$$
$$= argmax_{v_{j\in V}} p(a_1, \dots, a_{n,}|v_j)p(v_j)$$

Estimating $p(a_1, \dots, a_{n,}|v_j)$ directly from the data is difficult because too much data observations are needed. The Naïve Bayes method is based on the assumption that the attributes are independent given the class:

$$p(a_1, \dots, a_{n,}|v_j) = p(a_1|v_j)p(a_2|v_j) \dots p(a_n|v_j).$$

Under this assumption v^* becomes $v_{NB} = argmax_{v_{j\in V}} p(v_j)\Pi_i p(a_i|v_j)$.

Learning a Naïve Bayes thus requires to estimate $p(v_j)$ and $p(a_i|v_j)$ from the data.

Though the assumption of conditional independence is often violated, extensive experiments over many years show that the algorithm works well in practice (Blockeel, 2013, pp. 218).

This is because if the independence condition is violated then $p(a_1, \dots, a_{n,}|v_j) \neq p(a_1|v_j)p(a_2|v_j) \dots p(a_n|v_j)$. However, the prediction by the classifier still holds as long as $argmax_{v_{j\in V}} p(a_1|v_j)p(a_2|v_j) \dots p(a_n|v_j) = argmax_{v_{j\in V}} p(a_1, \dots, a_{n,}|v_j)$.

Finally, there is no single parameter for assessing the quality of a classifier. It depends on the application of the classifier. Different methods to evaluate classifiers are described in Blockeel (2013). The most common way to evaluate a classifier is its accuracy. Accuracy is the probability of making correct prediction. True positive rate (TP

rate) and false positive rate (FP rate) are also used to evaluate a classifier. TP rate is defined as the rate of true positives i.e. instances correctly classified as a given class. On the other hand, FP rate is defined as the rate of false positives i.e. instances falsely classified as a given class (Blockeel 2013). An alternative to accuracy-based evaluation is cost-based evaluation. This method distinguishes among mistakes that are worse than others. Depending on the context, misclassifying a positive instance as a negative can be worse than misclassifying a negative as positive. One popular cost-based evaluation in machine learning is the Receiver Operating Characteristics (ROC) diagram. The area under the ROC curve is another parameter of evaluation of classifier. DeLong et al. (1988) discusses about this approach. In this approach accuracy is measured by the area under the ROC curve. The area of one means perfect prediction and area of 0.5 means no predictive value (random guessing) (DeLong et al. 1988). Kappa statistics is another method to judge the quality of a classifier. It is a chance-corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement (Viera and Garrett 2005).

3. Methods

In this research Naïve Bayes algorithm is used to build the classifiers. First one is to predict positioning test score based on gender and academic information of the students in their high school education. In order to see the impact of gender similar classifier was built without using the gender attribute. The next one is to predict first year result of bachelor study based on positioning test score and previous academic information. The research also uses overall accuracy, kappa statistics, TP rate, FP rate, and ROC area to evaluate the efficiency of the classifiers. In order to construct and evaluate the classifiers, data mining software Waikato Environment for Knowledge Analysis (WEKA) is used. This is an open source software with a collection of machine learning algorithms for data mining tasks. These algorithms can be applied directly to a dataset or can be called from Java code. After loading the dataset WEKA provides the option to look at the data and preprocess it. It provides summary statistics for each attribute (mean, variance, attribute relationships, % of missing value). It also provides an alternative to explore the dataset using interactive visualization (Hall et al. 2009). The statistics and visual representation provided by WEKA help to identify potential problems and help to decide if an action needs to be taken (Hall et al. 2009).

4. Data Collection

While subscribing for the positioning test, participants provide the information about their gender and hours of mathematics conducted during high school education. The information about the mathematics score and total score during the high school education was collected from a questionnaire filled in during the positioning test. The score of this test is collected by the Tutorial Services of the Faculty of Engineering Science of KU Leuven. The result of first year of bachelor program is collected from the KU Leuven data bank. This dataset contains the information about 418 students regarding their gender, previous study related information (hours of math in the final years of high school education, percentage of math score and percentage of year total in the last year of high school education), score of the positioning test and corresponding results of the first year of the engineering bachelor study of the participants.

5. Results and Discussion

This section provides the result of Naïve Bayes based classifier for predicting the success and failure of the positioning test score based on the background information about mathematical achievement in the high school and gender. The result of another classifier which is built to predict the first year bachelor result is presented in order to see the usefulness of the positioning test.

5.1 Prediction of Positioning Test

The classifier is built to predict the positioning test score based on the gender and academic information of high school education. This section includes the exploration of the dataset from which the classifiers are trained, attribute selection, construction and evaluation of prediction models.

5.1.1 Exploring Dataset for Predicting Positioning Test Score

Four attributes (gender, hours of math, % of math and % of year total) are used in order to predict the positioning test score (grouped as pass and fail). The interpretation of the attributes and the groups of positioning test scores are presented in table 1.

Table 1: Description of the attributes for predicting positioning test score

Attribute	Description
Gender	Gender of the students
Hours of Math	Hours of Math conducted during the high school education (Typically there are two options for high school students to choose for mathematics course. One is 6 hours and the other is 8 hours.)
% of Math	Percentage of Math score obtained in the last year of their high school education
% of Year Total	Percentage of total score obtained in the last year of their high school education
Target Attribute	
Grouped Positioning Test Score	Positioning test score is labeled into two categories: Pass (10 or above in the scale of 20), Fail (0-9 in the scale of 20)

5.1.2 Attribute Selection and Construction of Classifier

Attribute selection is performed on all input data using information gain ranking filter in WEKA.

Table 2: Ranking of attributes according to information gain

Ranked Attributes	Information Gain
% of Math	0.0649
Hours of Math	0.0603
Gender	0.0493
% of Year Total	0.0402

The result of the information gain filtering (Table 2) shows that the percentage of math scores and hours of math during high school education have higher information gain for predicting the positioning test score in comparison to gender and percentage of total score obtained in the last year of their high school education. Then the classifiers are built using Naïve Bayes for predicting positioning test score (labeled as pass and fail) based on the previous information about the students (gender, % of math, hours of math conducted, % of total score) using 10-fold crossover validation.

5.1.3 Evaluation of Classifier

Correctly classified instances: 68.36%
Kappa statistic: 0.35

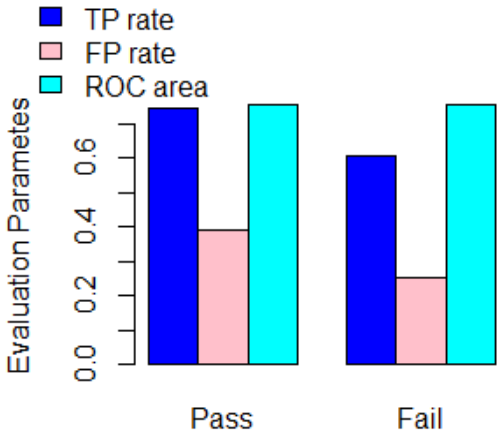

Figure 1: Class-wise (pass and fail) accuracy comparison of naïve bayes classifier

The percentage of correctly classified instances (accuracy) is 68.3%. It shows the percentage of test instances that are correctly classified. Kappa value 0.35 means that the classifier is doing better than chance (kappa value zero). Figure1 shows the class-wise comparison of different parameters (TP rate, FP rate and ROC area). The aim of the analysis is to see the predictive ability of the attributes (gender, previous mathematics experience etc.) on the positioning test score. The classifier provides similar result in classifying both groups. The ROC area is same (0.753) for both classes. The ROC area indicates that classifier is predicting better than random guessing (0.5) in order to predict the success and failure of positioning test based on the previous information about their secondary education and gender.

In order to see the impact of gender in predicting positioning test score, similar classifier (Naive Bayes based) is built without using the gender attribute and compared it with the classifier with gender attribute.

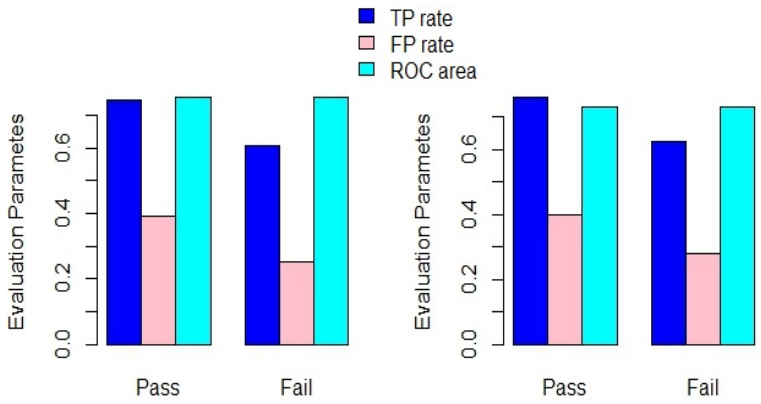

Figure 2: Comparison of naive bayes classifier with and without gender attribute

Figure 2 shows that there is a very little influence of gender on predicting positioning test score. The proportion of TP rate and FP rate and ROC area (0.73) in predicting both classes without gender attribute is close to that of with gender attribute (ROC area 0.753). Moreover, by exploring the data set it is observed that the number of male candidates participated in the positioning test and the number of them who have scored ten or above in the test is much higher than that of female candidates. Therefore, gender does not influence predicting positioning test score given a higher number of hours in mathematics. The students with better mathematics experience (higher hours of math, % of math) in their secondary education are more preapared for the engineering positioning test.

5.2 Prediction of First Year of Bachelor Result

Exploration of dataset, selection of attributes, construction, and evaluation of classifier for predicting first year result of bachelor study based on positioning test score and previous academic information is presented in this section.

5.2.1 Exploring Dataset for Predicting First Year of Bachelor Program

Along with the attributes used to predict positioning test score, grouped positioning test score and score of language test which was conducted for one group of engineering student during the mid of first semester is also used as an attribute for predicting first year result of bachelor study. The target attribute is labelled as A (No. of non-tolerable fail is zero), B (one non-tolerable fail), C (more than one non-tolerable fail). The term non-tolerable fail means the course in which a student has failed non-tolerably (i.e. scored 7 or lower in the scale of 20).

5.2.2 Attribute Selection and Construction of Classifier

The result (Table 3) of the information gain shows that the grouped positioning test score provides much higher information gain than those of other attributes.

Table 3: Ranking of attributes according to information gain

Ranked Attributes	Information Gain
Grouped Positioning Test Score	0.1486
% of Math	0.0649
Hours of Math	0.0603
Language Test	0.0581
Gender	0.0493
% of Year Total	0.0402

The classifier is built to predict the first year result of bachelor based on the positioning test score along with the previous information about the students. Similar model is built without using the positioning test score and compared the result with the previous model in order to analyze the usefulness of the test.

5.2.3 Evaluation of Classifier

Correctly classified instances: 61.02%
Kappa statistic: 0.3617

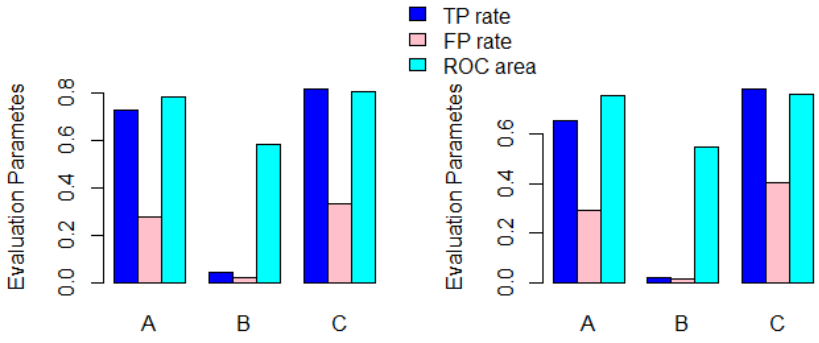

Figure 3: Class-wise accuracy comparison of naïve bayes classifier to predict first year bachelor result with all attributes (left) and without (right) positioning test score

The kappa value (0.3617) of the classifier (with all attributes) shows that it predicts better than guessing. Though the overall accuracy of the classifier is 61%, it provides much better result in predicting class A (group of students with no non tolerable fail) and C (group of students with more than one non tolerable fail) than B (group of students with one non tolerable fail) (Figure 3 (left)). Therfore, the classifier is well predictor of students who perform good (the

group of students with no non tolerable fail) and bad (the group of students with one non tolerable fail) of first year bachelor study based on their past academic information, gender, and positioning test score.

On the other hand, the overall accuracy decreases to 56% while predicting the study progress without the positioning test score. As the goal of the model is to predict the students who have less chance of doing well in engineering, the accuracy of class C is more important than that of class A. In comparing with the classwise accuracy of the two classifiers, it is seen that the model without the positioning test score predicts the students who will more likely have difficulty in their engineering education (TP rate of group C) as good as that of classifier with all attributes. However, wrong prediction of students who actually do not face difficulty (FP rate of group C) are more than the classifier with all attributes. Therefore, better balance of TP rate and FP rate is observed in the first classifier for group C (ROC area 0.8) than that of the second classifier (ROC area 0.75) (Figure 3).

6. Conclusion

The research provides the implementation of predictive learning algorithms in order to search for the engineering study success indicators. This implementation is achieved by Naïve Bayes based classifier. The first classifier is built to predict the test score based on the prior academic achievement and gender. The result shows that the classifier can predict both pass and fail of positioning test equally well. However, the contribution of gender is observed very low. Even without considering gender, the classifier provides similar accuracy in predicting pass and fail of the positioning test score. Therefore, the prior academic achievement, not the gender of the students influences the score of the positioning test score.

The second classifier is constructed to predict the first year result of bachelor based on the test score and prior academic achievement. For realizing the predictive power of the test score, similar classifier is also built without using this score. The comparison of both classifiers shows that the overall accuracy of the classifier is lower when they are trained without using the test score. Moreover, the accuracy of predicting students who have worst result (class C) is higher in case of with test score than without score. Therefore, the test score along with prior academic achievement is well predictor for the academic success of engineering. Moreover, this test score is better predictor for the students who are more likely to do poorly.

However in all the above cases there might be other factors that can influence the academic result (e.g. effort, motivation, study strategy, job demand, socio economic status, culture etc.) in addition with the score of the positioning test, mathematical experience, gender, and first year bachelor result. The classifiers are built based on the attributes that mainly contain the information about academic results. So in order to further improve the accuracy of the classifiers, attributes related to study strategy and technique can also be used.

Acknowledgement

The author gratefully acknowledges Professor Tinne De Laet, Faculty of Engineering Science, KU Leuven, for her guidance throughout the study and suggestions, comments, and practical support as a daily supervisor during the master's thesis. The author would also like to acknowledge Tutorial Services of the Faculty of Engineering Science of KU Leuven for providing the data to carry out the research.

References

ATTRACT Project, *Enhance the Attractiveness of Studies in Science and Technology*, Rep. no. 978-91-7501-127-1, Sweden, 2012.

Bouckaert, R. R., Bayesian network classifiers in weka for version 3-5-7, *Artificial Intelligence Tools*, *11*(3), 369-387, 2008.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., and Scuse, D., WEKA Manual for Version 3-6-0, University of Waikato, Hamilton, NewZealand, 2008.

Blockeel, H., Machine Learning and Inductive Inference. *Leuven: ACCO,* 2013.

CNES (Center for the New Economy and Society), The future of jobs report 2018, Geneva: World Economic Forum, 2018.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L., Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*, 837-845, 1988.

Duderstadt, J. J., Engineering for a changing world, *Holistic Engineering Education* (pp. 17-35), Springer New York, 2010.

EU Skills Panorama Analytical Highlight, Science, technology, engineering and mathematics (STEM) skills, 2012.

Hall, C., Dickerson, J., Batts, D., Kauffmann, P., & Bosse, M., Are we missing opportunities to encourage interest in STEM fields? *Journal of Technology Education Vol. 23 No. 1, Fall 2011, 32-45,* 2011.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H., The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18, 2009.

Mitchell, T. M., Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, *45*, 1997.

Moses, L., Hall, C., Wuensch, K., De Urquidi, K., Kauffmann, P., Swart, W., & Dixon, G., Are math readiness and personality predictive of first-year retention in engineering? *The Journal of psychology*, *145*(3), 229-245, 2011.

Skills Panorama, Researchers and engineers: skills opportunities and challenges, Retrieved from https://skillspanorama.cedefop.europa.eu/en/analytical_highlights/researchers-engineers-skills-opportunities-and-challenges-2019-update, 2019 update.

Skills Panorama, Researchers and engineers: skills opportunities and challenges, Retrieved from https://skillspanorama.cedefop.europa.eu/en/analytical_highlights/researchers-engineers-skills-opportunities-and-challenges-2016

Steenkamp, H., Nel A. L. and Carroll, L., Retention of engineering students, IEEE Global Engineering Education Conference (EDUCON), Athens, pp. 693-698, doi: 10.1109/EDUCON.2017.7942922, 2017.

UNESCO, Measuring gender equality in science and engineering: the SAGA Toolkit, *SAGA Working Paper* 2, Paris, 2017.

Vanderoost, J., Callens, R., Vandewalle, J. P. L., De Laet, T., *Diversity and the Academic Engineering Positioning Test in Flanders: Impact on Female Students and Students with Disabilities*. Proc. of 42nd Annual Conference Birmingham, UK, 2014.

Viera, A. J., & Garrett, J. M., Understanding inter observer agreement: the kappa statistic. *Fam Med*, *37*(5), 360-363, 2005.

Biography

Saeka Rahman is a lecturer in the department of Computer Science and Engineering at International Standard University, Dhaka, Bangladesh. She earned M.Sc. in Artificial Intelligence majoring Engineering and Computer Science from KU Leuven, Belgium which is among the top 50 universities of the world university ranking. Prior to this, Ms. Rahman completed her B.Sc. and M.Sc. in Applied Physics, Electronics and Communication Engineering from the University of Dhaka. Her current research interest is in the area of engineering education, artificial intelligence – particularly machine learning and data mining. She is a member of IEEE.