European Journal of Educational Research
Volume 11, Issue 1, 325 - 337.
ISSN: 2165-8714
https://www.eu-jer.com/

Supervised Learning Applied to Graduation Forecast of Industrial Engineering Students

Natalia Gil Canto
Federal University of Amazonas, BRAZIL

Marcelo Albuquerque de Oliveira[*]
Federal University of Amazonas, BRAZIL

Gabriela de Mattos Veroneze
Federal University of Amazonas, BRAZIL

Received: September 1, 2021 ▪ Revised: November 10, 2021 ▪ Accepted: December 1, 2021

Abstract: The article aims to develop a machine-learning algorithm that can predict student's graduation in the Industrial Engineering course at the Federal University of Amazonas based on their performance data. The methodology makes use of an information package of 364 students with an admission period between 2007 and 2019, considering characteristics that can affect directly or indirectly in the graduation of each one, being: type of high school, number of semesters taken, grade-point average, lockouts, dropouts and course terminations. The data treatment considered the manual removal of several characteristics that did not add value to the output of the algorithm, resulting in a package composed of 2184 instances. Thus, the logistic regression, MLP and XGBoost models developed and compared could predict a binary output of graduation or non-graduation to each student using 30% of the dataset to test and 70% to train, so that was possible to identify a relationship between the six attributes explored and achieve, with the best model, 94.15% of accuracy on its predictions.

Keywords: *Engineering retention, supervised learning, classification, graduation forecast.*

To cite this article: Canto, N. G., Oliveira, M. A., & Veroneze, G. M. (2022). Supervised learning applied to graduation forecast of industrial engineering students. *European Journal of Educational Research, 11*(1), 325-337. https://doi.org/10.12973/eu-jer.11.1.325

Introduction

The development of technologies using machine learning has shown explosive growth in the processes of creating products or services currently delivered to the market. This area of research emerges as a branch of artificial intelligence and exists as a basic principle of technologies aimed at speech recognition on smartphones, forecasting prices for the stock exchange, recommending films on streaming platforms, identifying diseases from the recognition of ultrasound images, among other applications, so that it can be described as "the science (and art) of programming computers in such a way that they can learn from data" (Géron, 2017).

Among the machine learning methods currently available, there is supervised and unsupervised learning. The first one will be the basis for this research because it involves learning from the visualization of labels (answers) of part of the data, so that the algorithm can predict new labels for the test data based on the learning relationships it found in the training data.

In addition, learning methods, in general terms, seeks to find answers to certain types of problems: those in which the relationships between the input data generate a continuous response (such as car prices), those in which this relationship generates a discrete response (such as vehicle types) or even those in which the answer is unknown so that it is necessary to search for a standardization by the algorithm according to the characteristics delivered to it.

In this context, a problem common to several undergraduate courses at public universities is identified: the low rate of student graduations. Veenstra et al. (2009) emphasizes that when it comes to sciences and engineering fields it turns out the retention rate is even lower, and the reasons can be cognitive, such as GPA or High School grades in general, non-cognitive, such as family support, financial difficulties, healthy or characteristics related to the color, gender, habits or expectations of each student.

So, delimited to the Industrial Engineering course at the Federal University of Amazonas (UFAM), the research development department, the statistics involved in this problem showed that, in a database of 364 students with an

[*] **Corresponding author:**
Marcelo Albuquerque de Oliveira, Federal University of Amazonas, Master in Production Engineering Program, Brazil.
✉ marcelooliveira@ufam.edu.br

admission period between 2007 and 2019, the index of graduations until the second semester of 2019 is 17.8% for the universe evaluated of possible graduates (considering the period of five years of course), in addition to presenting an index of 12.4% of retired or dropout students (considering the universe of evaluated samples).

Thus, the research starts from the hypothesis that certain parameters related to the performance of each student throughout the course can describe their result (graduation or non-graduation). Therefore, it aims to identify this binary output through a classifying machine learning algorithm that will learn the relationship between these data by visualizing them and predicting the results for data not previously seen.

Literature Review

The engineering retention problem

The problem involved in this work has been extensively addressed in recent decades, as for example in Lin et al. (2008), which seeks, through cognitive (GPA, grades in sciences or mathematics, among others) and non-cognitive (motivation, leadership, expectations, among others) characteristics, to estimate the retention of first-year students of engineering using neural networks, so that this work achieved about 78% probability of detecting student retention, but only 40% for non-retention, indicating that non-cognitive and cognitive characteristics could partially describe the problem, being necessary the combination of other variables that can also influence the persistence of these students, such as family, economics and health issues.

Similarly, using data from the University of Michigan, Veenstra et al. (2009) considers attributes similar to those of Lin et al. (2008), such as: High School Academic Achievement, Quantitative Skills, Study Habits, Commitment to Career and Educational Goals, Confidence in Quantitative Skills, Commitment to Enrolled College, Financial Needs, Family Support and Social Engagement. And it starts from the hypothesis that there is a direct correlation between the retention of engineering students and their GPA, so that, using logistic regression, it can identify that the characteristics High School Academic Achievement, Quantitative Skills, Commitment to Career and Education Goals and Confidence in Quantitative Skills predicted student success (GPA).

More recently, also seeking to understand the issues involved in retaining engineering students, French et al. (2021), works on the understanding both genders about the psychological cost involved in the course, given that the perception of students may differ when focused on minorities, such as the case of women in engineering. This work shows that there is a perception on the part of this minority group, but it is a group with greater probability of graduation, compared to male respondents.

Santacroce (2018) also emphasizes these results, showing that women are more likely to graduate with an engineering degree if they remain in the course after 2 years, even though their self-confidence is negatively affected by the stereotypes and majority men environment they face during the course (Jagacinski, 2013). Thus, the gender is an attribute to be considered as having great influence on the students results, as well as the skin color (Davis & Finelli, 2007; Green et al., 2019; Palmer et al., 2011; Reichert & Absher, 1997; Ye et al., 2021).

Besides that, a study carried out by Craig (2011) also emphasizes factors influencing retention and attrition of engineering students at historically black colleges and universities, such as: students working long hours brought on by insufficient financial aid; difficulty of the curriculum and poor teaching styles. In this study, the author also highlights possible strategies to solve these questions: enhanced advisement, tutorial, and mentoring activities.

Similarly, Fletcher and Anderson-Rowland (2000) also considered strategies related to mentoring and advisement to improve the performance of a group of sixteen engineering students and achieved amazing results, including an increase of 79% on the cumulative GPA of these students, indicating feasible ways to solve the retention problem, which were also extensively investigated over the years (Chelberg & Bosman, 2019; Desai & Stefanek, 2017; Hartman et al., 2019; Lisberg & Woods, 2018; Shahhosseini et al., 2020; Stromei, 2000).

Additionally, Honken and Ralston (2013) also discuss potential challenges that might be affecting the engineering low retention rate and highlights the lack of preparation in math and science as the top reason to students transfer out of engineering, followed by financial challenges and lack of time to study as the main reason to left university.

This last reason is interesting since engineering courses differ from other majors mainly by the programmatic activities, where the curriculum requires engineering students to not only participating in educational enriching activities, but also to gaining marketable experiences (Lichtenstein et al., 2010), which highly affects the daily time these students dedicate to study.

Furthermore, several other research sought to understand about the motivations behind the retention of students from Science, Technology, Engineering, and Mathematics (STEM) courses, considering different attributes, especially the cognitive ones, as addressed in this work (Coletti et al., 2014; García-Ros et al., 2019; Hieb et al., 2015; Koenig et al., 2012; Wang et al., 2015).

And comparing this common problem between STEM majors, Almatrafi et al. (2017) and Godfrey et al. (2010) found out even though the engineering retention is high and an issue to be worked on due to multiple reasons, like difficulties on understanding academic concepts, self-esteem or time issues, the persistence rates for this major is higher compared to college of sciences, highlighting the importance of studies related not only to engineering but, also, focused on the STEM science fields in general.

Logistic regression and XGBoost

First, logistic regression can be described as a linear model adequate when we have dichotomous outcome variables (Lemon et al., 2003; Subasi & Ercelebi, 2005). Kurt et al. (2008) emphasizes that this model is competent when predicting presence (or absence) of a characteristic or outcome based on values of predictor variables, a concept that allow us to understand the method as appropriate to find out the relationship between the characteristics involved in this research.

Second, XGBoost is a supervised learning algorithm based on gradient boosting decision trees (Dhaliwal et al., 2018). It's responsible to achieve state-of-the-art results in many machine learning competitions (Chen & Guestrin, 2016; Nielsen, 2016), especially those involving tabular data. Dhaliwal et al. (2018) and Chen and Guestrin (2016) also highlights that XGBoost is highly effective in reducing computing time by providing optimal use of memory resources, which is one of its great benefits since computational cost is still a challenge when dealing with neural networks solutions, that's why this algorithm has potential to be, also, an adequate solution to the graduation forecast problem presented.

Artificial Neural Networks - MLP

According to Akhgar et al. (2019) and Ngah and Bakar (2017), artificial neural networks, as the nomenclature suggests, are based on the human brain. This tool has been used as a prediction strategy in areas such as: maintenance management (Susto et al., 2015), identification of defective parts (Wang et al., 2019), diagnosis of diseases (Jiang et al., 2020), landslide predictions (Pham et al., 2017), among other applications. In a simplified way, this simulation created artificially, considers the information transfer process through a standardized scheme and composed of multiple processing units (Gehr et al., 2018). According to Tiwari and Khare (2015) the organization of these units takes place through layers that relate weights and inputs in order to identify the ideal outputs of the model.

The diagram below shows one of the primary artificial neuron models, called the Perceptron model (Géron, 2017), this model indicates the inputs of an artificial neuron (characteristics of a problem), its synaptic weights (with which the inputs will be combined in a linear transformation), the result of a summation of these combinations and an activation function responsible for performing a transformation, commonly, nonlinear.

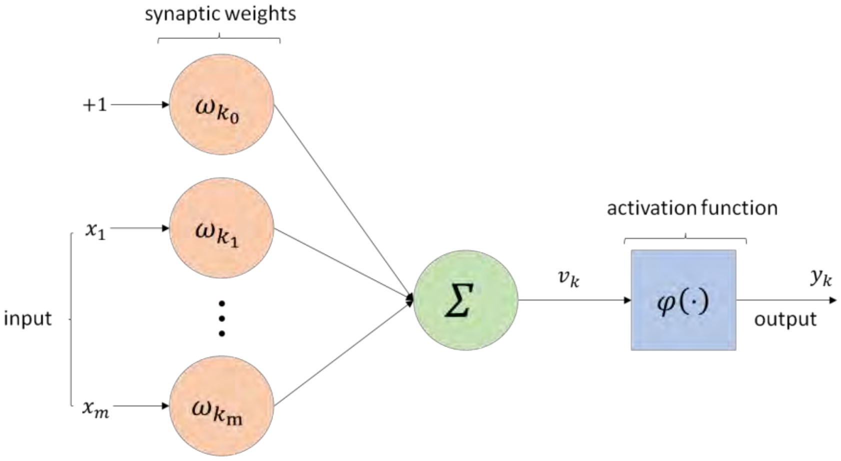

Figure 1. Perceptron Artificial Neuron Model

This model follows the vector representation below:

$$y_k = \varphi(\boldsymbol{w}_{\boldsymbol{k}}^T \boldsymbol{x})$$

On what $\boldsymbol{w}_{\boldsymbol{k}} = [w_{k0} \cdots w_{km}]^T$ and $\boldsymbol{x} = [1 \dots x_m]$. Thus, the model admits linear classification, being able to predict outputs that can be separated by means of a hyperplane (binary).

Activation Functions

The activation functions are used with a view to perform linear transformations or not of the output v_k. These functions are important in an artificial neural network because it limits the output of a neuron (Ngah et al., 2016), influencing the network's flexibility and, consequently, its efficiency.

Géron (2017) shows that, in principle, sigmoidal functions have long been used as activation functions, like: hyperbolic tangent, logistic function, among others; due to its satisfactory behavior in each layer of the network.

However, with the advent of deep learning techniques, the ReLu rectifier activation function (Rectified Linear Unit) was more cost-effective compared to its performance in approximations (weight adjustments) in the data training stage, in addition to reducing computational cost (Lomuscio & Maganti, 2017). Currently, there are several variations of this function.

The architecture of a neural network can be divided by two modes of propagation, namely: Feedforward Neural Network (FNN) and Recurrent Neural Network (RNN). On the one hand, for the first method, according to Jiang et al. al (2020), the network computes the direction between the layers in a direct way, in a single direction, until the exit of the network, following the equation below:

$$y(n) = \phi(x(n);\theta)$$

So that the network performs a mapping of the inputs $x(n)$ to an output $y(n)$ through the set of parameters θ. On the other hand, the second method, according to Wanto et al. (2017), involves feedback in the network that allows the model to recalculate its synaptic weights. This characteristic defines a recurring network as a network that has memory. So, as new information is fed, the network searches for past information that can characterize it more accurately. This model can be described by the equation below:

$$y(n) = \phi(x(n), y(n-1);\theta)$$

Where the mapping of entries $x(n)$ it is done considering the immediately previous output and its due adjustment parameters for the output calculation $y(n)$. This means we are dealing with a high flexible network depending on the number of layers and its depth, which can be a challenge when working with small datasets.

Methodology

Research Goal

The article aims to develop a machine-learning algorithm that can predict the graduation of students in the Industrial Engineering course at the Federal University of Amazonas based on their performance data.

Research Strategy

The data collection considered the performance report of students in the Industrial Engineering course available on the e-Campus portal.

Eligibility Criteria

The target audience of the research consists of evaluating the data of students from the Industrial Engineering Course at the Federal University of Amazonas.

Sample and Data Collection

The data collection and analysis methodology considered the performance report of students in the Industrial Engineering course available on the e-Campus portal. This report was generated based on the enrollment period between 2007 and 2019, and consists of information from 364 students. These characteristics were then evaluated based on their possible influence on the characteristic of interest in this work: the student's graduation.

Thus, the analysis methodology, as well as the supervised learning models, were based on the development of an algorithm in Python programming language under version 3.8.5 on Google Colaboratory (code available on GitHub repository SL-Algorithms) using 3 different methods: Logistic regression, a shallow MLP and the decision tree-based XGBoost, aiming to compare and choose the best of them. The construction of the prediction models followed the flow shown in the Figure 2.

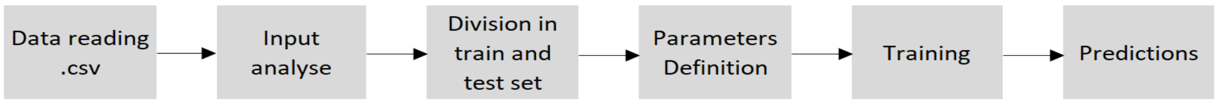

Figure 2. Flow of the Algorithm Construction

Firstly, the conversion of the input attributes file to the appropriate reading format was considered, which is done by .csv files (separated by commas). After that, the complete data were plotted using graphs that describe each of the attributes separately, in addition to the presentation of the correlation between attribute-attribute and attribute-output, as previously described. As a result, the data were randomly divided into two parts: training and testing. So that 30% of the data was destined for the testing stage, which has the objective of reproduce an *on-campus* application and cannot be used during the training stage to avoid any bias on the predictions.

Thus, before submitting the data to the model, it was necessary to assess their balance, considering that the number of graduated students is lower than the number of non-graduated students and this situation can lead to a biased learning and impact the model's performance on the test set. The Figure 3 explains this situation.

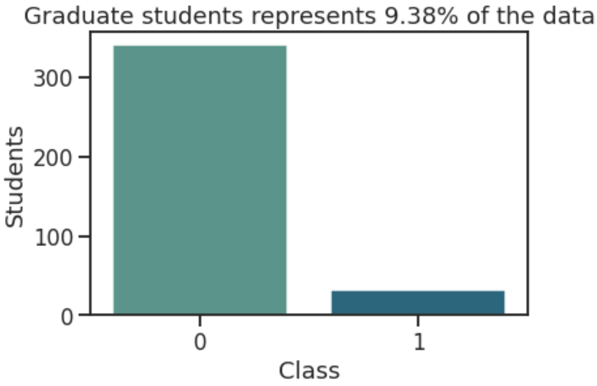

Figure 3. Unbalanced Data

As can be seen, only 9.4% of the data represent graduate students (class 1), which creates difficulties for the model to learn the characteristics that lead to this class. Thus, the application of a resampling feature in the training set was necessary, in order to artificially increase the amounts of data whose output is equal to 1 (graduate students). This process can be seen in Figure 4, which indicates a balance between the two outputs.

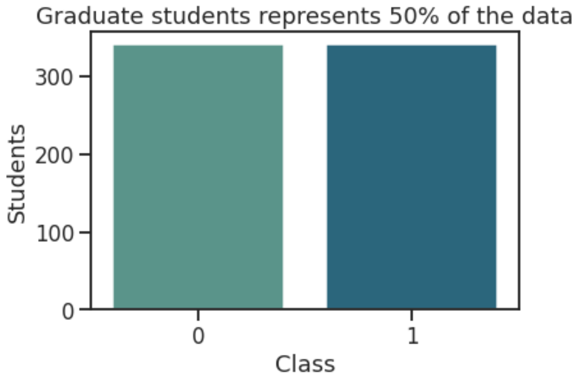

Figure 4. Balanced Data

With the data balanced and the initial parameters defined, the models were trained and the predictions along with the test data were performed for the Logistic Regression, MLP and XGBoost models.

Analyzing of Data

The analysis of the characteristics of entry considered the following attributes: grade-point average, semesters taken, courses terminations, type of school in high school, lockouts and dropouts, as shown in Figure 5 and Figure 6.

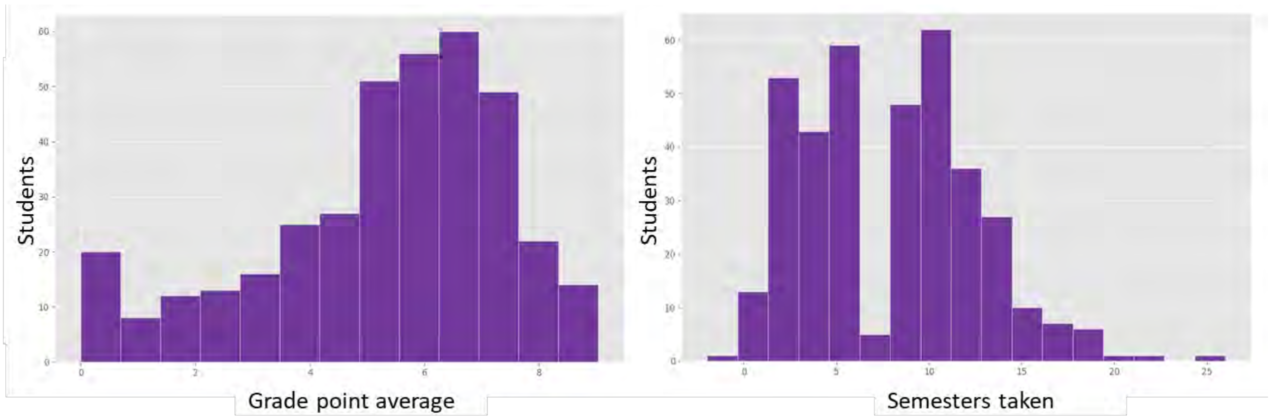

Figure 5. Input Characteristics

The attributes related to the grade-point average and the semesters taken were divided into histograms, so that it is possible to notice, in the first case, a distribution displaced, since the average for approval considered at the university is equal to 5.0. In the second, it is possible to note two distinct peaks, the first at point 5 and the second at point 10, indicating that the data does not show uniform numbers of students per admission period.

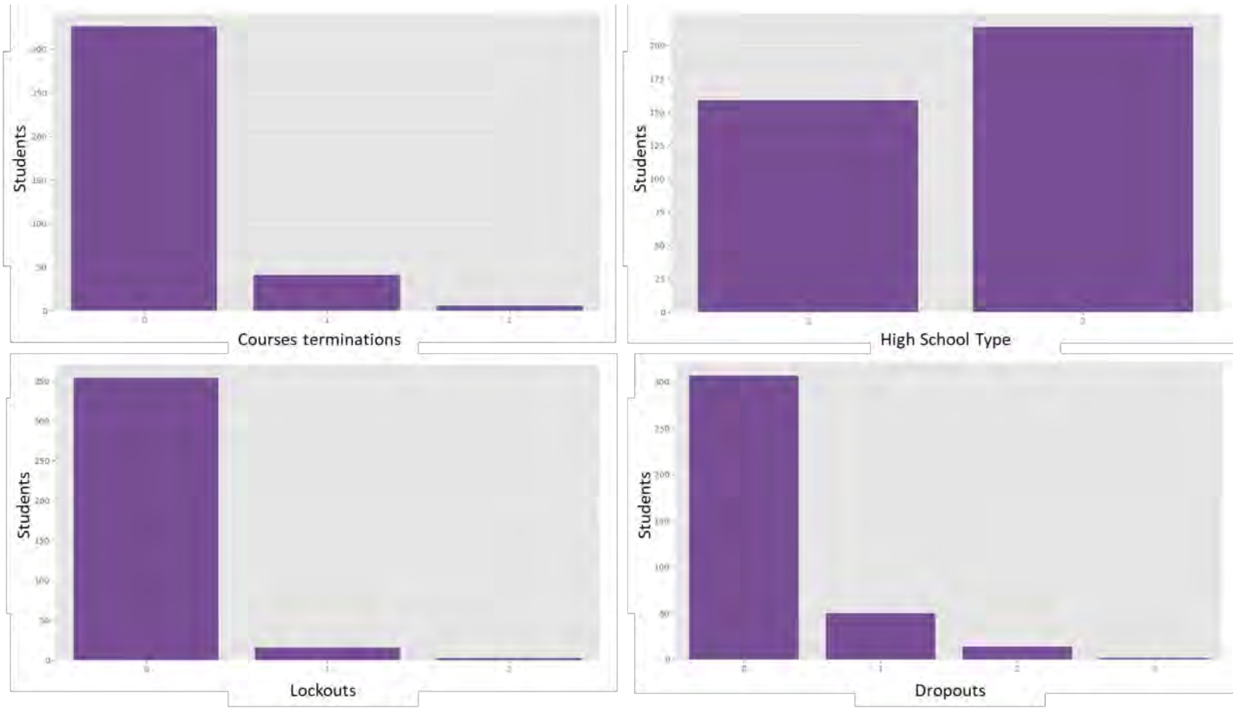

Figure 6. Input Characteristics

In contrast, the attributes that are divided into classes could be presented in the graphs above. Firstly, the number of courses terminations have a high rate of retired students, an even greater number than graduates, as will be seen later, whether in the current course or not. On the side, the graph was divided into only two classes, namely: private schools (type 1) and public schools (type 2), so that the number of students from type two schools are more expressive.

Then, the graph that lists the number of students by the number of enrollment lockouts indicates a low index for it, which differs from the graph on the side, which lists dropouts from the course and has a high index. It is important to highlight that the enrollment lockouts and dropouts from the course are attributes that can be performed more than once by the same student, since there is a period for canceling the action in the university system. Thus, the correlation between the attributes and the data output could be assessed, as can be seen in the Figure 7.

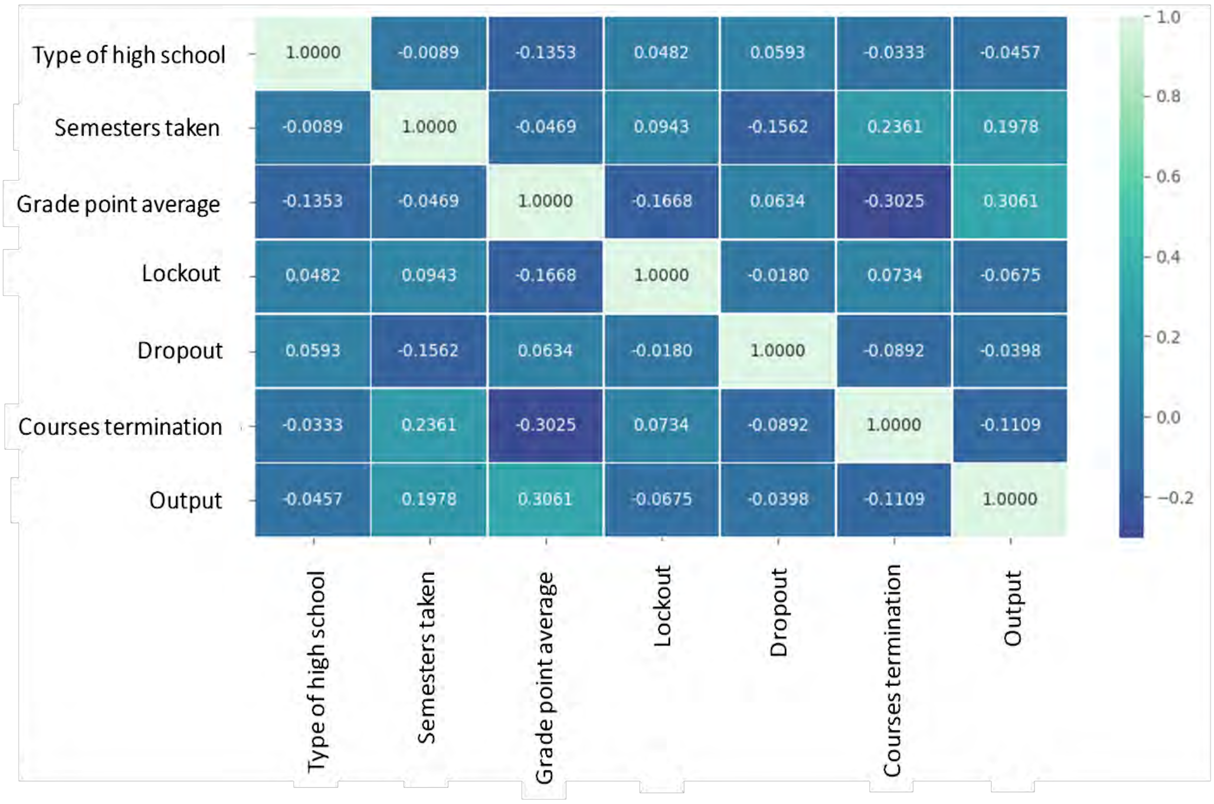

Figure 7. Correlation Map

The attribute-attribute and attribute-output correlation map can identify the influence between all characteristics and their outputs, even if represented by classes. Thus, in the map above, it is possible to identify correlation levels close to zero in most of the points, however, positive and negative correlations, which indicate an influence between the data, could also be seen. For example, the relationship between the grade-point average and the number of courses terminations, which is inversely proportional, and the relationship between the grade-point average and the output, which appears to be in direct proportion.

Findings / Results

First, the training set was analyzed using logistic regression, as this is a categorical problem, and this analysis allowed us to verify an adequate accuracy value, given the limited amount of data for analysis. Then, the data were trained through a shallow MLP, with only two intermediate layers containing 6 and 5 neurons, respectively, using ReLU as activation function and Adam as it's optimizer. In this case, the model had a performance similar to that found through logistic regression, however, it presented a very bumpy error surface (many local minima). Finally, the data were processed using XGBoost, which achieved an accuracy of around 94.15%, higher than the previous ones.

Table 1. Models' Accuracy Comparison

Model	Main parameters	Accuracy	Incorrect classifications
Logistic Regression	Penalty: l2 Solver: lbfgs	91.71%	17
MLP	2 layers (6-5 neurons) Initial learning rate: 0.01 Epochs: 500 Batch size: 32	91.22%	18
XGBoost	Max depth: 3 Eta: 0.3 Epochs: 1000	94.15%	12

Therefore, considering XGBoost as the best estimator, its confusion matrix can be plotted to analyze the behavior of this classifier in the test set, which simulates the *on-campus* operation of the proposed model.

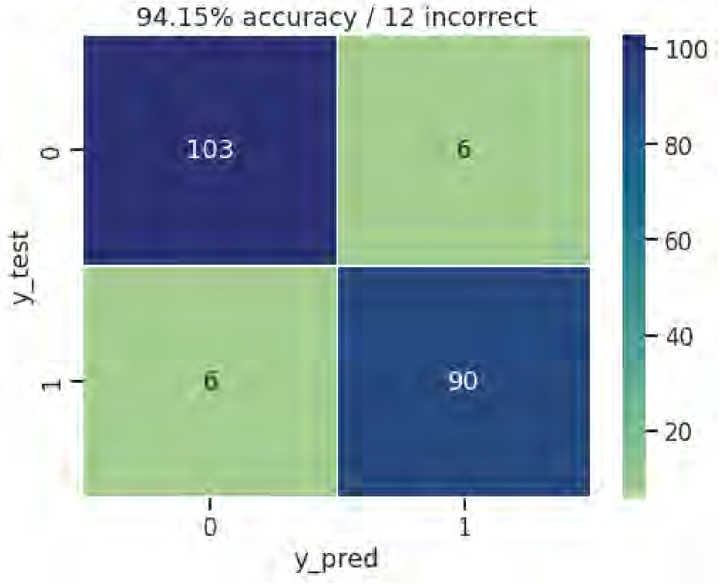

Figure 8. Confusion Matrix for XGBoost

As can be seen, the matrix above lists the number of errors and hits in the model (y_pred) in relation to the test data (y_test), so that, on the one hand, the main diagonal indicates the correctness rates of the model, which is around 103 true negatives and 90 true positives. On the other hand, the secondary diagonal shows the quantities of incorrect classifications performed by the model, being, then, 6 false positives and 6 false negative. With that, some metrics could be calculated, as can be seen bellow, which indicates the performance of the model for the test set.

Table 2. Model Evaluation

Metric	Formulation	Results
Accuracy	$\text{acc}(\hat{y}(x)) = \frac{TP + TN}{TP + TN + FP + FN}$	94.15%
Precision	$\text{Precision}(\hat{y}(x)) = \frac{TP}{TP + FP}$	93.75%
Recall	$\text{Recall}(\hat{y}(x)) = \frac{TP}{TP + FN}$	93.75%
F-measure (1.0)	$F_m = \frac{(m+1) \times Recall(\hat{y}(x)) \times Precision(\hat{y}(x))}{Recall(\hat{y}(x)) + m \times Precision(\hat{y}(x))}$	93.75%
Error rate	$p_e(\hat{y}(x)) = \frac{FP + FN}{TP + TN + FP + FN}$	5.85%
Specificity	$\text{Specificity}(\hat{y}(x)) = \frac{TN}{TN + FP}$	94.50%
False negative rate	$p_e^+(\hat{y}(x)) = \frac{FN}{TP + FN}$	6.25%
False positive rate	$p_e^-(\hat{y}(x)) = \frac{FP}{TN + FP}$	5.50%

The first metric is accuracy, which relates the number of standards classified correctly in relation to the entire set of test data, indicating a result around 94% (12 incorrect classifications). This metric is one of the most important for the analysis of the model. However, other characteristics need to be evaluated in parallel, so the robustness of the model is verified.

In this way, then, precision is presented, a metric that corresponds to the proportion of data referring to the positive class correctly classified in relation to all the examples attributed to the positive class. As can be seen, the precision obtained a good percentage of 93.75%, close to the accuracy achieved.

In addition, the recall was also calculated and obtained a result of 93.75%, which indicates the rate of true positive, in a way that relates the proportion of examples of the positive class correctly classified and presented a result according to the precision, since their number of false negatives and false positives are equal.

In this way, the metric f-measure could be calculated, which relates the two metrics mentioned earlier through a weighted harmonic mean, where 'm' is the weighting factor. This weighting factor is responsible for the importance given to each of the metrics involved in the calculation, so that the value of 'm' used was equal to 1.0, which indicates that the importance of recall and precision are the same. Thus, the metric obtained, as expected, 93.75%, indicating a value close to 100%, which is the ideal value of the f-measure.

Then, the model's error rate can also be seen in the table above, a metric that is inverse to the accuracy presented a result around of 5.85%. So, specificity was also calculated, indicating the rate of true negative, one of the best results among all the calculated metrics, which was around 94.50%, showing how the class with real values, in face of the artificially created values of class 1, represents gain to the model even though the data is balanced.

Finally, the metrics involving the false negative and false-positive rates were calculated, so that the differences in classification could be viewed more intensely. Thus, the first was around 6.25% compared to 5.50% presented by the second, a great performance considering the set size and the number of attributes used.

Discussion

Through the comparative analysis between the supervised learning methods presented, it was possible to observe, firstly through accuracy, that logistic regression proved to be effective in the given classification task, although it was not the model with the highest performance. This result is interesting given that the method has a simple, linear proposal, used in older researches, such as Veenstra et al. (2009), and more recent ones, such as French et al. (2021), confirming it's importance as a competitive estimator and which, in addition, was able to present a superior result than the one found through the more flexible, fully connected neural network, indicating the data used in training, to some extent, can be separated by means of a straight line.

Furthermore, the results presented by the MLP network could show that, even though we have greater flexibility and can learn through synaptic adjustments of the network weights, this learning can be impacted by the amount of data available, so that the error surface can present numerous local minimums and not allowing an adequate search for the solution that brings the highest performance in the test set.

Finally, we highlight that the performance of the decision tree-based algorithm, XGBoost, could confirm the statement made by Chen and Guestrin (2016) and Nielsen (2016) about its ability to generate state-of-the-art results in various machine learning problems, especially those involving tabular data, such as the one presented in this work. As seen, the algorithm was placed as an intermediate solution from the point of view of flexibility, that is, it is not as rigid as logistic regression, nor as flexible as the neural network, generating the most adequate solution for the binary classification problem we have.

In this way, supervised learning techniques could identify the existence of a relationship between the six characteristics that were worked on and the data output: grade-point average, semesters taken, courses terminations, high school type, lockouts and dropouts. And this allows us to state that these attributes have a strong influence on student results and can be able to describe the problem presented quite adequately, highlighting the potential of the estimator in face of several works that seek to predict, in a certain way, the graduation or retention of students until the end of their courses, as in the surveys mentioned previously: Lin et al. (2008), Veenstra et al. (2009) and French et al. (2021).

Furthermore, the existence of a relationship between these attributes makes possible for us to discuss about the ideal characteristics to invest on with strategies to increase student's graduation rate having a small universe of possibilities (a total of 6 possibilities).

In other words, we can analyze, mainly, the characteristics that could be improved by the university, such as: grade-point average, investing on different methodologies and developing a research to understand from the students and teachers what could be improved or maintained; type of high school, investing on extra courses to improve basic lacks on math or science subjects; courses terminations, lockouts and dropouts, implementing scholarships, assistance programs and upgrading laboratories and common areas to possibility a higher academic time dedication as well as the development of practical skills inside the university.

With that, we can highlight the semesters taken is a fix attribute used to indicate to the learning algorithm the engineering course has a mean time to be completed, so, for now, the main strategies can be applied in 5 of the 6 attributes used on this research and the implementation order can be made by considering the correlation between each of them and the output of interest (students graduation or non-graduation).

Conclusion

The development of a machine-learning model through supervised learning algorithms is part of the range of predictive methodologies created under the aspect of artificial intelligence. This possibility is based primarily on the development of computer systems that are capable of storing a large amount of data.

This application started from the hypothesis that several factors are responsible for the graduation (or not) of students of the Industrial Engineering course at the Federal University of Amazonas and as it could be seen, together, certain selected attributes had an influence on the study output.

More than the methodology used in the creation of a supervised learning algorithm that could learn based on this dataset, the work could present to the academic community a new way of applying information available in the university's own system, indicating the use of these data not only to statistical studies, but to studies that are aligned with 4.0 industry, such as artificial intelligence, and that bring the novelty of a rising area that needs to be further explored by the Industrial Engineer based on the multidisciplinary intrinsic to this profession.

Thus, the way the machine learning area has been developing in recent years indicates that the implementation of a learning model is not limited to the areas of engineering or computer science, and can currently be applied to various problems, including problems common to the educational area, as seen on previous research, were the authors explored the reasons or motivations regarding to engineering students retention by using different techniques but achieving similar results, which implies we already have ways to understand this problem and propose effective solutions.

With that, the good performance of the developed algorithm, as discussed in the previous session, indicates the importance of using these methods within the university itself as a base methodology for implementing improvements within the Industrial Engineering department or, even, the others that can make use of this algorithm for performance analysis.

Thus, according to what was discussed, one of the main characteristics that are related to the output of the model is the data of grade-point average, with an influence proportional to the output. Thus, the study of teaching methodologies used in each of the disciplines is important for the individual student's performance, in such a way that this performance is an important answer in relation to psychological aspects such as: student interest, excitement with the course, encouragement from teachers and self-esteem within the department. As a result, the department's investment in methods that can reduce dropouts, lockouts and course terminations is essential to the student's success, since the improvement in the grade-point average becomes a consequence of these factors mentioned because they are somewhat correlated.

In addition, as it can be seen in the present work, the improvements to the model will occur according to the number of examples both classes (0 and 1) grow, as the model will be able to improve their learning with more examples and also with a greater number of students in class 1 (trained), since there would be a natural data balance.

With that, the Industrial Engineering department will also be able to make use of the same algorithm developed in this work. Besides, it will be able to increase the database to improve the model predictions and check, then, the impacts happening in the department arising from new strategies implemented.

Recommendations

Although this research carried out an analysis of the performance of students from the Industrial Engineering Course during the years 2007 to 2019, it is recommended that the same analysis be carried out for the entire period of activity of the course, from its institutional creation to the present day. In addition, a similar study can be carried out within the Faculty of Technology and the courses it offers to the community, enabling analysis by course and the development of an institutional strategic plan.

Also, it would be interesting to collect data of the final grades of students in the department in each of the subjects offered, so that the model output is based on the hypothesis that each of the subjects has a direct influence on student success or failure in the course, which would indicate a better direction of the department towards specific improvements to each of them.

Limitations

As limitations, it can be highlighted that this study involved the analysis only of students from the Industrial Engineering course at the Faculty of Technology, among other courses in the academic unit. The study was also limited to analyzing the performance of students from the year 2007 and not since the creation of the course, which dates back to 2004. Besides that, there were limitations related to the data treatment, since they were not on a format suitable to the algorithm, which makes important to the interested teachers or researchers to first open a project with a selected team only destinated to the analysis and organization of these data to later use.

Acknowledgements

The authors thank the Production Engineering Department (DEP) of the Federal University of Amazonas (UFAM) for their support and structure. Special thanks to the Master's Program in Production Engineering at the Federal University of Amazonas for the support of its professors for the research carried out.

This work was conducted in partnership with the Federal University of Amazonas (UFAM) within the scope of the Master's Program in Production Engineering (PPGEP). The authors thank the support obtained from Fundação de Apoio à Pesquisa do Estado do Amazonas (FAPEAM) for facilitating and finance the research under RESOLUÇÃO N. 006/2020 - POSGRAD 2020.

Authorship Contribution Statement

Canto: Conceptualization, design, analysis, writing. de Oliveira: editing/reviewing, supervision, critical revision of manuscript, statistical analysis, securing funding. de Mattos Veroneze: editing/reviewing, supervision, critical revision of manuscript, final approval.

References

Akhgar, A., Toghraie, D., Sina, N., & Afrand, M. (2019). Developing dissimilar artificial neural networks (ANNs) to prediction the thermal conductivity of MWCNT-TiO2/Water-ethylene glycol hybrid nanofluid. *Powder Technology, 355,* 602-610. https://doi.org/10.1016/j.powtec.2019.07.086

Almatrafi, O., Johri, A., Rangwala, H., & Lester, J. (2017, June 24-28). *Board 65: Retention and persistence among STEM students: a comparison of direct admit and transfer students across engineering and science* [Paper presentation]. 2014 ASEE Annual Conference & Exposition, Columbus, Ohio, United States. https://doi.org/10.18260/1-2--27899

Chelberg, K. L., & Bosman, L. B. (2019). The role of faculty mentoring in improving retention and completion rates for historically underrepresented STEM students. *International Journal of Higher Education*, *8*(2), 39-48. https://doi.org/10.5430/ijhe.v8n2p39

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In B. Krishnapuram & M. Shah, (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Coletti, K. B., Wisniewski, E. O., Shapiro, R. L., DiMilla, P. A., Reisberg, R., & Covert, M. (2014, June 15-18). *Correlating freshman engineers' performance in a general chemistry course to their use of supplemental instruction* [Paper presentation]. 2014 ASEE Annual Conference & Exposition, Indianapolis, Indiana, United States. https://doi.org/10.18260/1-2--20214

Craig, W. O. (2011). Strategies for improving the retention of engineering and technology students at historically black colleges and universities HBCU. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies*, *2*(5), 561-570. https://tuengr.com/V02/561-568.pdf

Davis, C.-S. G., & Finelli, C. J. (2007). Diversity and retention in engineering. *New Directions for Teaching and Learning, 111*, 63-71. https://doi.org/10.1002/tl.287

Desai, N., & Stefanek, G. (2017, March 2-5). *A literature review of the different approaches that have been implemented to increase retention in engineering programs across the United States* [Paper presentation]. 2017 ASEE Zone II Conference, San Juan, Puerto Rico. http://zone2.asee.org/papers/proceedings/3/117.pdf

Dhaliwal, S. S., Nahid, A.-A., & Abbas, R. (2018). Effective intrusion detection system using XGBoost. *Information, 9*(7). https://doi.org/10.3390/info9070149

Fletcher, S. L., & Anderson-Rowland, M. (2000, June 18-21). *Developing retention strategies for women that promote student success in engineering and the applied sciences* [Paper presentation]. 2000 Annual Conference, St. Louis, Missouri, United States. https://doi.org/10.18260/1-2--8284

French, S. E., Bego, C. R., Hieb, J. L., & Ralston, P. A. (2021, July 26-29). *Psychological cost, gender, and retention among engineering students* [Paper presentation]. 2021 ASEE Virtual Annual Conference Content Access, Virtual Conference. https://peer.asee.org/37622

García-Ros, R., Pérez-González, F., Cavas-Martínez, F., & Tomás, J. M. (2019). Effects of pre-college variables and first-year engineering students' experiences on academic achievement and retention: a structural model. *International Journal of Technology and Design Education, 29*(4), 915-928. https://doi.org/10.1007/s10798-018-9466-z

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., & Vechev, M. (2018). *AI2: Safety and robustness certification of neural networks with abstract interpretation* [Paper presentation]. IEEE Symposium on Security and Privacy (SP), San Francisco, CA, United States. https://doi.org/10.1109/SP.2018.00058

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools and techniques to build intelligent systems* (1st ed.). O'Reilly Media.

Godfrey, E., Aubrey, T., & King, R. (2010). Who leaves and who stays? Retention and attrition in engineering education. *Engineering Education*, *5*(2), 26-40. https://doi.org/10.11120/ened.2010.05020026

Green, C., Dika, S., & Smith, A. (2019, June 15-19). *Board 137: Persistence of women of color in undergraduate engineering programs* [Paper presentation]. ASEE 2019 Annual Conference and Exposition, Tampa, Florida, United States. https://doi.org/10.18260/1-2--32248

Hartman, H., Forin, T., Sukumaran, B., Farrell, S., Bhavsar, P., Jahan, K., Dusseau, R., Bruckerhoff, T., Cole, P., Lezotte S., Zeppilli, D., Macey, D. (2019). Strategies for improving diversity and inclusion in an engineering department. *Journal* *of Professional Issues in Engineering Education and Practice*, *145*(2). https://doi.org/10.1061/(ASCE)EI.1943-5541.0000404

Hieb, J. L., Lyle, K. B., Ralston, P. A., & Chariker, J. (2015). Predicting performance in a first engineering calculus course: Implications for interventions. *International Journal of Mathematical Education in Science and Technology, 46*(1), 40-55. https://doi.org/10.1080/0020739X.2014.936976

Honken, N. B., & Ralston, P. (2013). Freshman engineering retention: A holistic look. *Journal of STEM Education: Innovations and Research, 14*(2), 29-37. https://bit.ly/3D0WvQz

Jagacinski, C. M. (2013). Women engineering students: Competence perceptions and achievement goals in the freshman engineering course. *Sex Roles, 69*(11-12), 644-657. https://doi.org/10.1007/s11199-013-0325-9

Jiang, J., Wang, H., Xie, J., Guo, X., Guan, Y., & Yu, Q. (2020). Medical knowledge embedding based on recursive neural network for multi-disease diagnosis. *Artificial Intelligence in Medicine, 103*, 1-12. https://doi.org/10.1016/j.artmed.2019.101772

Koenig, K., Schen, M., Edwards, M., & Bao, L. (2012). Addressing STEM retention through a scientific thought and methods course. *Journal of College Science Teaching, 41*(4), 23-29.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications, 34*(1), 366-374. https://doi.org/10.1016/j.eswa.2006.09.004

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine, 26*(3), 172-181. https://doi.org/10.1207/S15324796ABM2603_02

Lichtenstein, G., McCormick, A. C., Sheppard, S. D., & Puma, J. (2010). Comparing the undergraduate experience of engineers to all other majors: Significant differences are programmatic. *Journal of Engineering Education, 99*(4), 305-317. https://doi.org/10.1002/j.2168-9830.2010.tb01065.x

Lin, J.-J., Malyscheff, A., & Imbrie, P. (2008, June 22-25). *Artificial intelligence methods to forecast engineering students' retention based on cognitive and non-cognitive factors* [Paper presentation]. 2008 Annual Conference & Exposition, Pittsburgh, Pennsylvania. https://doi.org/10.18260/1-2--4315

Lisberg, A., & Woods, B. (2018). Mentorship, mindset and learning strategies: An integrative approach to increasing underrepresented minority student retention in a STEM undergraduate program. *Journal of STEM Education*, *19*(3), 14-20. https://www.learntechlib.org/p/184625/

Lomuscio, A., & Maganti, L. (2017). *An approach to reachability analysis for feed-forward Relu neural networks.* https://arxiv.org/abs/1706.07351

Ngah, S., & Bakar, R. A. (2017). Sigmoid function implementation using the unequal segmentation of differential lookup table and second order nonlinear function. *Journal of Telecommunication, Electronic and Computer Engineering, 9*, 103-108. https://jtec.utem.edu.my/jtec/article/view/2637

Ngah, S., Bakar, R. A., Embong, A., & Razali, S. (2016). Two-steps implementation of sigmoid function for artificial neural network in Field Programmable Gate Array. *ARPN Journal of Engineering and Applied Sciences, 11*(7), 4882-4888. https://bit.ly/3nScpH6

Nielsen, D. (2016). *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* [Master's thesis, Norwegian University of Science and Technology]. NTNU Open. http://hdl.handle.net/11250/2433761

Palmer, R. T., Maramba, D. C., & Dancy, T. E. (2011). A qualitative investigation of factors promoting the retention and persistence of students of color in STEM. *The Journal of Negro Education, 80*(4), 491-504. http://www.jstor.org/stable/41341155

Pham, B. T., Bui, D. T., Prakash, I., & Dholakia, M. B. (2017). Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA 149*(1), 52-63. https://doi.org/10.1016/j.catena.2016.09.007

Reichert, M., & Absher, M. (1997). Taking another look at educating African American engineers: The importance of undergraduate retention. *Journal of Engineering Education, 86*(3), 241-253. https://doi.org/10.1002/j.2168-9830.1997.tb00291.x

Santacroce, A. (2018). Determining strategies for the Embry-Riddle Aeronautical University college of engineering faculty to use to increase the retention rate of women in their undergraduate engineering programs. *The Compass, 1*(5), 30-36. https://scholarworks.arcadia.edu/thecompass/vol1/iss5/6

Shahhosseini, A. M., Moayed, F. A., & McLeod, A. (2020, June 22-26). *Strategies for increasing enrollment, retention, and graduation in two baccalaureate degree STEM programs: Mechanical Engineering Technology (MET) and Safety Management (SM)* [Paper presentation]. 2020 ASEE Virtual Annual Conference Content Access, Virtual Conference. https://monolith.asee.org/public/conferences/172/papers/29080/view

Stromei, L. K. (2000). Increasing retention and success through mentoring. *New Directions for Community Colleges*, *2000*(112), 55-62. https://doi.org/10.1002/cc.11205

Subasi, A., & Ercelebi, E. (2005). Classification of EEG signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine, 78*(2), 87-99. https://doi.org/10.1016/j.cmpb.2004.10.009

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: a multiple classifiers approach. *IEEE Transactions on Industrial Informatics 11*(3), 812-820. https://doi.org/10.1109/TII.2014.2349359

Tiwari, V., & Khare, N. (2015). Hardware implementation of neural network with Sigmoidal activation functions using CORDIC. *Microprocessors and Microsystems, 39*(6), 373-381. https://doi.org/10.1016/j.micpro.2015.05.012

Veenstra, C. P., Dey, E. L., & Herrin, G. D. (2009). A Model for freshman engineering retention. *Advances in Engineering Education, 1*(3), 1-33. https://files.eric.ed.gov/fulltext/EJ1076050.pdf

Wang, J., Fu, P., & Gao, R. X. (2019). Machine vision intelligence for product defect inspection based on deep learning and Hough transform. *Journal of Manufacturing Systems, 51,* 52-60. https://doi.org/10.1016/j.jmsy.2019.03.002

Wang, M.-T., Degol, J., & Ye, F. (2015). Math achievement is important, but task values are critical, too: examining the intellectual and motivational factors leading to gender disparities in STEM careers. *Frontiers in Psychology, 6*, 1-9. https://doi.org/10.3389/fpsyg.2015.00036

Wanto, A., Windarto, A. P., Hartama, D., & Parlina, I. (2017). Use of binary sigmoid function and linear identity in artificial neural networks for forecasting population density. *International Journal of Information System & Technology, 1*(1), 43-54. https://doi.org/10.30645/ijistech.v1i1.6

Ye, N., Fok, T. Y., Collofello, J., & Coronella, T. (2021, July 26-29). *Common and uncommon characteristics of engineering student retention after the first year in university* [Paper presentation]. 2021 ASEE Virtual Annual Conference Content Access, Virtual Conference. https://doi.org/10.18260/1-2--36813