Predicting Whether Student will continue to Attend College or not using Deep Learning

Qasem M. M. Zarandah¹, Samy S. Abu-Naser², Moshera M. Elgohary³, Zakaria K. D. AlKayyali¹, Bassem S. Abu-Nasser¹, Ashraf M. Taha¹,
¹Ph.D. Student, University Malaysia of Computer Science & Engineering (UNIMY), Cyberjaya, Malaysia.
²Department of Information Technology, Faculty of Engineering and Information Technology, Al-Azhar University, Gaza, Palestine
³Faculty of Computers and Information Menofia University, Cairo, Egypt

Corresponding author: Prof. Dr. Samy S. Abu-Naser, abunaser@alazhar.edu.ps

Abstract: According to the literature review, there is much room for improvement of college student retention. The aim of this research is to evaluate the possibility of using deep and machine learning algorithms to predict whether students continue to attend college or will stop attending college. In this research a feature assessment is done on the dataset available from Kaggle depository. The performance of 20 learning supervised machine learning algorithms and one deep learning algorithm is evaluated. The algorithms are trained using 11 features from 1000 records of previous student registrations that have been enrolled in college. The best performing classifier after tuning the parameters was NuSVC. It achieved Accuracy (91.00%), Precision (91.00%), Recall (91.00%), F1-score (91.00%), and time required for training and testing (0.04 second). Additionally, the proposed DL algorithm scored: Accuracy (93.00%), Precision (93.00%), Recall (93.00%), F1-score (93.00%), time required for training and testing (0.66 second) for predicting whether student will continue to attend college or not.

Keywords: Deep Learning, Attend College, Student, Prediction

1. Introduction

Initially, Neural Networks (NN) were widely known as Artificial Neural Networks (ANN). A structure like the implementation of biological neurons, with a learning structure based on probability that receives input data to decide on an output. Input data includes texts, images and audio files. Deep Learning utilizes artificial neurons to serve high dimensional data. Deep Learning can perform tasks involving communication pattern and information processing.

Figure 1 illustrates the basic structure of NN, with the first layer named as the input and the last as the output with the middle layers being the hidden layers hence the name “deep”. With further research, structure was improved with the implementation of backpropagation. The backpropagation algorithm has two parts, the first is for the input to forward the data towards the output, the second is to evaluate and estimate the error and backpropagate error values to the neural networks for error correction.

![Figure 1: Structure of Neural Networks](image)

Each node is a feature extracted from the portion before. The number of neurons on each layer vary for each system, as the number of neuron on each layer must be able to capture the essential layer. The first layer models are generally high in number of neurons.

Deep learning sets a whole new area of research with the architecture that has fixed many issues that were in the traditional systems where the model was used for specific usage or particular commands. DL were able to be a more uniform model that would be applied on many applications as well as different users. The architecture of DL can be implemented on several applications, such as the Google Assistant that can be used on mobile phones (Android), also can be used in search engines (Google.com) as well as being implemented on household devices (Google Home), and also used in video captioning such as the video captioning on YouTube.
In this paper, some data classification techniques are applied to the evaluation of the dataset to predict whether student will continue going to college or not by finding the best classification technique in accordance with these measures: Accuracy, F1-score, Recall, Precision and time performance.

We used 20 machine learning algorithms to predict whether student will continue going to college or not. Furthermore, a deep learning model was proposed for the same purpose.

2. Research Questions

The aim of this study is to answer the following questions:

- Can we use the dataset at hand to create a supervised machine learning classifier to predict reliably whether a student go to college or not?
- Can we use the dataset at hand to create a supervised Deep learning classifier to predict reliably whether a student go to college or not?
- Would the performance of Deep Learning techniques is better than the machine learning techniques in this case?

3. Machine Learning Algorithms

Machine learning algorithms can be applied to solve multiple problems. Classification can be used to assign a category to items or answer “yes” or “no” questions. An example of a classifier is a program that categorizes news articles or a spam filter that answers the question “spam” or “not spam”. Regression algorithms can be used to predict values for items such as housing prices. Ranking can be used to order items according to a criterion and clustering can be used to partition items into homogeneous regions [30].

A decision tree is an example of a machine learning algorithm. Decision trees contain a root with a question. The root has branches to more nodes with questions and each path of nodes and branches ends with a leaf containing a label. The tree is traversed by taking decisions at each node based on the question until a new value is found and the label for the leaf is returned as the answer to the root question [30, 41].

Ensemble learning methods are boosting algorithms that combine several weak learners into a single strong learner. One example of a boosting algorithm, AdaBoost [30], builds this strong learner from several weak learners by making the next learner focus on improving the mistakes of the previous learner. The learners are assigned weights by the algorithm after each iteration and the weights are used to merge the weak learners into a strong learner.

Linear models predict values by seeking a hypothesis of a hyperplane that has the smallest outcome of an error function. The linear model has to predict the value of y for the input value x. Using a set of known inputs x and labeled output values y a model is evaluated which has the smallest error, in this case distance of the line from the points in the graph. The model with the smallest error is represented as a line passing through the points. New input values x can then have their output values y predicted using the model by inspecting where the line meets the input value x on the y axis. Linear models such as Logistic Regression can also be used to solve classification problems [13, 30].

Support Vector Machines (SVM) can be applied to both regression and classification problems. SVMs seek to find a hyperplane which has the maximum margin from all inputs in the training set. SVMs differentiate and improve upon Linear Models with a better error or loss function. SVMs can be even more improved upon with the use of kernels to define non-linear decision boundaries [18, 30].

In addition to the aforementioned algorithms a great deal of other algorithms exist. k-Nearest-Neighbor consider neighboring objects in the dataset to train a better model [10]. Naive Bayes (NB) algorithms use Bayes’s theorem to compute conditional probabilities [20]. Discriminant Analysis classifies objects in a dataset by identifying the best feature to discriminate between classes [25].

4. Methodology

In this section we give in details the data collection, preparation, feature analysis, data splitting, modeling the proposed deep learning algorithm, training, validating and testing all the algorithms used in this study.

4.1 Dataset

The dataset was collected from Kaggle depository. The dataset consist of 11 features and contains 1000 records. Table 1 lists all the features available from the dataset.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-school</td>
<td>object</td>
<td>Academic or Vocational</td>
</tr>
<tr>
<td>School-accreditation</td>
<td>object</td>
<td>A or B</td>
</tr>
<tr>
<td>Interest</td>
<td>object</td>
<td>Very Interested, Uncertain, Less Interested, Quiet Interested, Not Interested</td>
</tr>
</tbody>
</table>
### 4.2 Feature Analysis

#### 4.2.1 Correlation Matrix

The highest in negative correlation will be parents-was-in-college – parent-age, and with positive is parent-salary – in-college, house-area – in-college and average-grades – in-college as illustrated in Figure 2.

![Correlation Matrix](image)

**Figure 2: Correlation Matrix**

#### 4.2.2 Histogram for distribution

**Figure 3** shows the distribution of the features: parent-age, parent-salary, house-area, and average-grades.

![Histogram for distribution](image)

**Figure 3: Histogram for distribution**

#### 4.2.3 Density plot for distribution

**Figure 4** illustrates the density plot distribution for parent-age, parent-salary, house-area, and average-grades. The overall distribution looks a little bit normal.
4.2.4 Outliers
Figure 5 does not show so many outliers, so my decision is to keep it without further cleaning.

4.2.5 Residence count based on college status
Overall the Urban is more likely to go to college and the opposite happens in Rural as shown in Figure 6.

4.2.6 Residence with parent salary based on college status
Figure 7 illustrates that the higher the salary the more likely to go to college for both Urban and Rural.
4.2.7 Interest count based on college status

From Figure 8, Very Interested and Uncertain Interest are the two most categories attended the college, and Less Interested is the most category who does not attend the college.

4.2.8 School type count based on college status

Figure 9 show that the Academic is the most category attended the college.

4.2.9 Gender count based on college status

From Figure 10, females are slightly higher to attend the college than Males.
4.2.10 Parent age, residence and parent–was-in-college:

Figure 11 illustrates that Urban parents was slightly higher to attend the college than Rural parents.

4.2.11 Parent salary, residence and parent–was-in-college:

Figure 12 show that Rural parents was higher salary and not attended to college and the opposite for Urban parents.

4.2.12 Parent salary, residence and in-college:

The more salary in both Rural and Urban parents the more likely for their child to attend the college as can be seen if Figure 13.
4.2.13 House-area, residence and in-college:
The more house area in both Rural and Urban the more likely for child to attend the college, that’s means by logic the family has more salary overall as can be seen if Figure 14.

4.2.14 Average-grades, gender and in-college:
The more average grades in both males and females the more likely to attend college as shown if Figure 15.

4.2.15 Average-grades and in-college:
The higher on average grades the more likely to attend the college as in Figure 16.
4.2.16 Parent-age and in-college:
Not so much effect by the age of parents for a child to attend the college as in Figure 17.

4.2.17 Parent- salary and in-college:
From Figure 18, the higher on parent's salary the more likely for the child to attend college.

4.2.18 House-area and in-college:
From Figure 19, it can be seen that the higher on house area the more likely for the child to attend college.

4.2.19 Converting categorical column to numeric ones
We have converted the categorical type features (type-school, school-accreditation, gender, interest, residence, parent-was-in-college, in-college) to numeric values.
4.2.20 Class (in-college) Distributions
We checked whether the class (in-college) is balanced or not. It turned out to be balanced as shown in Figure 20.

![Class (in-college) Distribution](image)

**Figure 20:** Class (in-college) distribution

4.3 Dataset splitting
We have split the current dataset into three datasets: Training, validating, and testing datasets. The ratio of splitting was (80%, 10%, 10%).

4.4 Description of the Algorithms Used in this Study
There are many algorithms of Machine Learning that can be applied for the prediction of whether the student continue to attend college or not. We have trained, validated and tested 20 various ML algorithms on our current dataset. The algorithms that were used for prediction and analysis are from different categories of machine learning algorithms to predict whether student will continue to attend college or not.

Furthermore, a deep learning model was proposed to predict whether student will continue to attend college or not. The proposed DL model consists of 6 Dense layers: one input layer (10 features), 4 hidden layers (128, 64, 32, and 16 neurons), and one output layer with 2 classes and softmax function as shown in Figure 21.

![DL Model Structure](image)

**Figure 21:** Structure of the proposed DL model
5. RESULT AND ANALYSIS

5.1 Performance Evaluation
We used the most popular performance measures for machine and deep learning algorithms such as: precision, recall, accuracy, and f1-score as outlined in eq. 1, 2, 3, and 4.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (1)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (2)
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)
\]

\[
\text{F1 – Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)
\]

Where TP: True Positives, TN: True Negatives, FP: False Positive, and FN: False Negatives

5.2 Performance of Used Algorithms
We used a group of machine learning and one proposed deep learning algorithms for the prediction if student will continue to attend college or not. The machine learning algorithms belong to different categories of machine learning like Naïve based, SVM, KNN, trees, analysis, and others. The proposed deep learning algorithm is custom made for the prediction if student will continue to attend college or not.

We have split the dataset into Training, Validating, and Testing. We have trained every algorithm individually using our training dataset, tested it and we made record of its accuracy, recall, f1-score, precision, and time need for training and testing.

Furthermore, we have trained the proposed deep learning algorithm using the same training dataset and cross-validated it using the validating dataset. We kept training the proposed DL algorithm until no room for improvement. We made record of the DL algorithm accuracy, recall, f1-score, precision, and time need for training, validating and testing. Part of the DL algorithm training, validating accuracies and losses are shown in Figure 23.
It turned out that the best Machine Learning algorithm was (NuSVC) for predicting whether student will continue to attend college or not. NuSVC model achieved Accuracy (91.00%), Precision (91.00%), Recall (91.00%), F1_score(91.00%), time required for training and testing (0.04 second) as can be seen in Table 2. Additionally, the proposed DL algorithm scored: Accuracy (93.00%), Precision (93.00%), Recall (93.00%), F1_score(93.00%), time required for training and testing (0.66 second) for predicting whether student will continue to attend college or not as can be seen in Table 3.

Table 2: Performance of the Machine Learning Algorithms

<table>
<thead>
<tr>
<th>Machine Learning Model-Name</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1_score</th>
<th>Time-in-Sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>NuSVC</td>
<td>91.00%</td>
<td>91.00%</td>
<td>91.00%</td>
<td>91.00%</td>
<td>0.04</td>
</tr>
<tr>
<td>GradientBoostingClassifier</td>
<td>89.50%</td>
<td>89.50%</td>
<td>89.50%</td>
<td>89.50%</td>
<td>0.15</td>
</tr>
<tr>
<td>MLPClassifier</td>
<td>89.00%</td>
<td>89.00%</td>
<td>89.00%</td>
<td>89.00%</td>
<td>0.80</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>88.50%</td>
<td>88.51%</td>
<td>88.50%</td>
<td>88.50%</td>
<td>0.22</td>
</tr>
<tr>
<td>LogisticRegressionCV</td>
<td>88.00%</td>
<td>88.00%</td>
<td>88.00%</td>
<td>88.00%</td>
<td>0.33</td>
</tr>
<tr>
<td>LGBMClassifier</td>
<td>87.50%</td>
<td>87.50%</td>
<td>87.50%</td>
<td>87.50%</td>
<td>0.11</td>
</tr>
<tr>
<td>QuadraticDiscriminantAnalysis</td>
<td>87.50%</td>
<td>87.50%</td>
<td>87.50%</td>
<td>87.50%</td>
<td>0.03</td>
</tr>
<tr>
<td>LogisticRegression</td>
<td>87.00%</td>
<td>87.00%</td>
<td>87.00%</td>
<td>87.00%</td>
<td>0.02</td>
</tr>
<tr>
<td>CalibratedClassifierCV</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>0.06</td>
</tr>
<tr>
<td>BaggingClassifier</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>0.06</td>
</tr>
<tr>
<td>LinearSVC</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>0.01</td>
</tr>
<tr>
<td>GaussianNB</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>0.01</td>
</tr>
<tr>
<td>LinearDiscriminantAnalysis</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>86.50%</td>
<td>0.02</td>
</tr>
<tr>
<td>AdaBoostClassifier</td>
<td>86.00%</td>
<td>86.00%</td>
<td>86.00%</td>
<td>86.00%</td>
<td>0.11</td>
</tr>
<tr>
<td>Perceptron</td>
<td>85.00%</td>
<td>85.00%</td>
<td>85.00%</td>
<td>85.00%</td>
<td>0.01</td>
</tr>
<tr>
<td>KNeighborsClassifier</td>
<td>84.00%</td>
<td>84.00%</td>
<td>84.00%</td>
<td>84.00%</td>
<td>0.01</td>
</tr>
<tr>
<td>LabelPropagation</td>
<td>83.50%</td>
<td>83.50%</td>
<td>83.50%</td>
<td>83.50%</td>
<td>0.04</td>
</tr>
<tr>
<td>DecisionTreeClassifier</td>
<td>83.50%</td>
<td>83.50%</td>
<td>83.50%</td>
<td>83.50%</td>
<td>0.01</td>
</tr>
<tr>
<td>ExtraTreeClassifier</td>
<td>79.50%</td>
<td>79.50%</td>
<td>79.50%</td>
<td>79.50%</td>
<td>0.01</td>
</tr>
<tr>
<td>NearestCentroid</td>
<td>79.00%</td>
<td>79.00%</td>
<td>79.00%</td>
<td>79.00%</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3: Performance of the Proposed Deep Learning Algorithm

<table>
<thead>
<tr>
<th>Deep Learning Model-Name</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1_score</th>
<th>Time-in-Sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed DL Model</td>
<td>93.00%</td>
<td>93.00%</td>
<td>93.00%</td>
<td>93.00%</td>
<td>0.66</td>
</tr>
</tbody>
</table>

The first aim of this study was to answer the following question: Can we use the dataset at hand to create a supervised machine learning classifier to predict reliably whether a student go to college or not? The answer for this question is yes, we were able to create a supervised machine learning algorithm and F1-score accuracy was 91%.

The second aim of this study was: Can we use the dataset at hand to create a supervised Deep learning classifier to predict reliably whether a student go to college or not?

The answer to this question was also yes and achieved F1-score of 93%.

The last aim of the study was: Would the performance of Deep Learning techniques is better than the machine learning techniques in this case?

The answer is yes, because the deep learning algorithm scored 93% in F1-score while the machine learning algorithm achieved 91%.

6. Conclusion

In this study, we used twenty Machine Learning algorithms and a deep learning algorithm for predicting whether student will continue to attend college or not. The dataset was collected from Kaggle Repository. In order to predict whether student will continue to attend college or not, a group of twenty machine learning and one deep learning algorithm were used. Among the machine learning models used, the best machine-learning algorithm was (NuSVC) for predicting whether student will continue to attend college or not. NuSVC model achieved Accuracy (91.00%), Precision (91.00%), Recall (91.00%), F1-score (91.00%), time required for training and testing (0.04 second). Additionally, the proposed DL algorithm scored: Accuracy (93.00%), Precision (93.00%), Recall (93.00%), F1-score (93.00%), time required for training and testing (0.66 second) for predicting whether student will continue to attend college.