*Application of Data Science and Machine Learning in the Prediction of College Dropout: A Data-Driven Predictive Approach*

Axel Frederick Félix Jiménez

Engineering in Computer Systems, National Polytechnic Institute

Zacatecas, México

frederick.felix@outlook.com

Isaul Ibarra Belmonte

Software Engineer, Center for Research in Mathematics

Zacatecas, México

Isaul.ibarra@cimat.com

Vania Stephany Sanchéz Lee

Engineering in Computer Systems, National Polytechnic Institute

Zacatecas, México

s.vanialee@gmail.com

Ezra Federico Parra González

Computer Science Department, Center for Research in Mathematics

Zacatecas, México

ezra.parra@cimat.com

Abstract — This paper presents a study on the prediction of student graduation or failure using two predictive models: K-Nearest Neighbors (KNN) and a forward sequential Artificial Neural Network (ANN). The models were implemented using a carefully selected set of independent variables. Relevant data were fed into these models and their efficiency was evaluated using metrics such as precision, accuracy and other performance measures.

The results obtained revealed that both the KNN model and the sequential forward ANN model achieved high efficiency in predicting student graduation or failure. The KNN model achieved an accuracy of 0.87% with an optimal value of K=3, while the sequential forward ANN model obtained an accuracy of 0.89% after 50 training epochs.

The efficiency achieved by these models is relevant in the context of predicting student academic outcomes. These models can provide a valuable tool to identify early on students at risk of not graduating and to take preventive measures to improve their academic performance.

In addition, the results obtained in this study were compared with other related studies in the field. Although there are differences in the variables used and the specific contexts, a general consistency in the results was observed, which supports the relevance and reliability of the predictive models used.

Keywords - data science, machine learning, education, mexico, improvement.

1. Introduction

In Mexico, the implementation of data science and machine learning in education has been limited compared to other countries. A scarcity of adoption and application of these technologies has been identified, as mentioned in a previous study. Only three articles were found that specifically addressed this topic in the Mexican context [1]. Challenges facing the implementation of data science and machine learning in education include lack of technological infrastructure, shortage of subject matter experts, and regulatory and data privacy barriers. Despite these challenges, projects and case studies have been identified that show the potential of these technologies in education in Mexico, such as the use of the GeoGebra application in teaching Binomial Probability and the evaluation of student perception during the COVID-19 pandemic. To boost the implementation of data science and machine learning in education, collaboration between government, educational institutions, the technology industry and civil society is required. This would involve training more data science experts, developing appropriate technology infrastructure, and creating clear policies and regulatory frameworks [2]. The successful application of these technologies in education in Mexico has the potential to improve educational quality and equity in the country.

1. Objectives

The main objective of this paper is to use machine learning and data science techniques to predict whether a college student will graduate or fail college. Predicting students' academic performance is an important challenge in education, as it can provide valuable information to identify and provide early support to those students who are at risk of dropping out or having difficulties in their studies.

The implementation of machine learning and data science in this context makes it possible to analyze large volumes of student data and use machine learning algorithms to identify relevant patterns and correlations. These predictive models can take into account a variety of variables, such as previous academic performance, class attendance, participation in extracurricular activities, and other individual student characteristics.

By predicting students' academic success or failure, more effective educational strategies and policies can be developed to improve student retention and promote academic success.[3] In addition, implementing these data-driven approaches can help educational institutions allocate resources more efficiently and provide personalized interventions to at-risk students.

In this article, we aim to explore the potential of applying machine learning and data science in predicting college dropout, as well as to identify the key factors that influence students' academic performance. In addition, we seek to evaluate the effectiveness of the predictive models developed and provide relevant information to improve educational quality and equity in the university context. The following is a list of specific objectives that will guide the development of this study and contribute to the advancement of the field of data-driven education:

1. Apply machine learning and data science techniques to predict the academic performance of university students.
2. Develop a predictive model that allows early identification of those students with a higher risk of dropping out or failing college.
3. Identify the key factors and variables that influence students' academic success or failure.
4. To evaluate the efficacy and accuracy of the predictive model in predicting student graduation or failure.
5. To establish a methodological and analytical framework that can be used in future research and applications related to the prediction of college dropout.
6. Demonstrate the potential and relevance of the application of machine learning and data science in education for informed, evidence-based decision making.
7. Contribute to the advancement and development of the field of data-driven education by fostering the implementation of technological and analytical solutions in educational institutions.

The purpose of using a specific dataset in this study is to collect relevant and representative data to analyze and predict students' academic success. The selected dataset provides detailed information on variables such as previous academic performance, attendance to previous grades in the first two semesters, occupation of parents (mother and father) and other factors that may influence student performance.

The methodology used to analyze the dataset and make the predictions is based on machine learning and data science techniques. Classification and regression algorithms are used to identify patterns and relationships in the data, and training and validation techniques are applied to ensure the accuracy and generalization of the predictive models.

1. Methodology
   1. Description of the Dataset used
      * 1. Characteristics of the dataset

The dataset used in this study contains 4424 records with 35 attributes. Each record represents an individual student and can be used to compare the performance of different algorithms for solving the same type of problem and for training in the area of machine learning. The attributes are grouped into five classes: demographic data, socioeconomic data, macroeconomic data, academic data at enrollment, and academic data at the end of the first and second semesters. The dataset is available in CSV format encoded in UTF8 and contains no missing values. The detailed description of each attribute can be found in Table 1. The data used in this study were obtained from internal and external sources of the institution, including the Academic Management System (AMS), the Teaching Activity Support System (PAE), annual data from the Directorate General of Higher Education (DGES), and the Contemporary Portugal Database (PORDATA).[4]

1. Attributes used grouped by class of attribute. It describes the aspects of each element in the dataset, grouped by classes of attributes depending on the type of attribute

| Class of attribute | Attribute |
| --- | --- |
| Demographic data | Marital status  Nationality  Displaced  Gender  Age at enrollment  International |
| Socioeconomic data | Mother’s qualification  Father’s qualification  Mother’s occupation  Father’s occupation  Educational special needs  Debtor  Tuition fees up to date  Scholarship holder |
| Macroeconomic data | Unemployment rate  Inlfation rate  GDP |
| Academic data at enrollmen | Application mode  Application order  Course  Daytime/evening attendance  Previous qualification |
| Academic data at the end of 1st semester | Curricular units 1st sem (credited)  Curricular units 1st sem (enrolled)  Curricular units 1st sem (evaluations)  Curricular units 1st sem (approved)  Curricular units 1st sem (grade)  Curricular units 1st sem (without evaluations) |
| Academic data at the end of 2nd semester | Curricular units 2nd sem (credited)  Curricular units 2nd sem (enrolled)  Curricular units 2nd sem (evaluations)  Curricular units 2nd sem (approved)  Curricular units 2nd sem (grade)  Curricular units 2nd sem (without evaluations) |
| Target | Target |

The dataset used in this study contains information on 17 undergraduate programs in different fields of knowledge, such as agronomy, design, education, nursing, journalism, management, social service and technologies. The CSV file containing the dataset is encoded in UTF8, which ensures compatibility with different systems and applications.[5] The CSV file containing the dataset is encoded in UTF8, which ensures compatibility with different systems and applications.[5]

* + - 1. Number Variables

The data set used in this study contains a total of 35 variables or attributes. These variables are grouped into different classes, which include:

Demographic data: these attributes provide information on demographic characteristics of the students, such as gender, age, nationality, etc.

Socioeconomic data: These attributes provide information on the socioeconomic background of the students, such as parents' level of education, type of housing, etc.

Macroeconomic data: These attributes contain information on macroeconomic indicators, such as unemployment rate, inflation rate and GDP.

Academic data at enrollment: These attributes provide information about the application and admission process for students, such as application mode, application order, chosen course, etc.

Academic data at the end of the first and second semester: These attributes contain information about students' academic performance at the end of the first and second semester, such as curricular units passed, grades obtained, etc.

* + - 1. SOURCE DATA

The data used in this dataset comes from a variety of sources, both internal and external to the institution. The data sources used include:

Academic Management System (AMS): this internal system of the institution provides data related to the academic management of students, such as enrollment information, academic progress, etc.

Teaching Activity Support System (PAE): This system developed internally at the institution provides data related to teaching activity, such as student attendance, grades, etc.

General Directorate of Higher Education (DGES): This external source provides annual data on admission through the National Competition for Access to Higher Education (CNAES).

Contemporary Portugal Database (PORDATA): This external source provides macroeconomic data on Portugal, such as unemployment rate, inflation rate, GDP, etc.

These data sources were used to collect demographic, socioeconomic, macroeconomic and academic information from the students [6].

* 1. Data Cleaning Process using RStudio

During the dataset preparation and cleaning process, several transformations were carried out to ensure data quality and consistency. One of the modifications made was the conversion of categorical variables to numerical. In particular, the case of the "Target" variable was addressed, which is the variable that is the focus of our study, since this variable indicates whether the student is “GRADUATE”, “ENROLLED” Y “DROPOUT” , it should be noted that this is our predictor variable..

To facilitate subsequent analysis and modeling, it was decided to assign numerical values to each of these categories. In this sense, it was established that "GRADUATE" corresponded to the value 1, while "DROPOUT" was assigned as 0. On the other hand, the category "ENROLLED" was eliminated from the dataset because it was not relevant for the objectives of the study.

These modifications made it possible to establish a consistent and adequate numerical representation for the "Target" variable, which is fundamental in the context of predicting students' academic success. It should be noted that all the transformations and changes made during the preparation and cleaning of the dataset were carried out using the RStudio environment, applying rigorous techniques and processes to guarantee the integrity of the data and obtain reliable results in the subsequent analysis.

* 1. Correlation and multiple linear regression method for variable selection.
     + 1. Correlation matrix calculation

A correlation matrix is a statistical tool that presents the correlation coefficients between sets of variables. The essence of a correlation matrix lies in its ability to succinctly showcase the correlation, a measure of the relationship, between two variables. Each cell within the matrix represents the unique relationship between two variables [7]. It is a table where each row and column of the table corresponds to a variable in a dataset. The intersection of a row and a column, that is, each cell in the table, contains the correlation coefficient between the two variables. This coefficient is a statistical measure that defines the degree to which these two variables move in relation to each other.

Before analyzing the relationships between the variables and the target variable, it is important to note that variables with more stars in the correlation matrix indicate a stronger correlation with the target variable. The correlation calculated between the variables based on the Target objective function is presented in a table below.

1. correlation with the target variable. with the correlation function in the target variable, the variables that have the most influence with the card variable were reflected.

|  |  |
| --- | --- |
| **Variable** | **Correlation level** |
| Maritalstatus |  |
| Applicationmode | \* |
| Applicationorder |  |
| Cours | \*\*\* |
| Daytimeeveningattendance |  |
| Previousqualification | \*\* |
| Nacionality | \* |
| Mothersqualification |  |
| Fathersqualification |  |
| Mothersoccupation | \*\* |
| Fathersoccupation |  |
| Displaced |  |
| Educationalspecialneeds |  |
| Debtor | \*\*\* |
| Tuitionfeesuptodate | \*\*\* |
| Gender | \*\*\* |
| Scholarshipholder | \*\*\* |
| Ageatenrollment | \*\* |
| International | \*\* |
| Curricularunits1stsemCredited | \*\*\* |
| Curricularunits1stsemEnrolled |  |
| Curricularunits1stsemEvaluations | \* |
| Curricularunits1stsemApproved | \*\*\* |
| Curricularunits1stsemGrade | \*\* |
| Curricularunits1stsemWithoutEvaluations |  |
| Curricularunits2ndsemCredited | \*\*\* |
| Curricularunits2ndsemEnrolled | \*\*\* |
| Curricularunits2ndsemEvaluations | \*\*\* |
| Curricularunits2ndsemApproved | \*\*\* |
| Curricularunits2ndsemGrade |  |
| Curricularunits2ndsemWithoutEvaluations |  |
| Unemploymentrate |  |
| Inflationrate |  |
| GDP |  | |

The table shows the correlations between the variables and the target variable (graduate-dropout-enrolled). Stars (\*) indicate the strength of the correlation, where more stars represent a stronger correlation. Variables with stronger correlations include "Cours", "Debtor", "Tuitionfeesuptodate", "Gender", "Curricularunits1stsemCredited","Curricularunits2ndsemCredited", "Curricularunits2ndsemEvaluations", and "Curricularunits2ndsemApproved". Variables with a dot (.) have a weak or non-significant correlation with the target variable..

* + - 1. Variable selection

The selection of variables with 2 and 3 stars in the correlation matrix is justified by their higher level of correlation with the target variable. By considering only these variables, we focus on those that present a stronger relationship with the outcome we are analyzing, be it graduation, dropout, or enrollment.

1. variable selection. based on the correlation function, the variables with the greatest influence on the target variable were selected.

|  |  |
| --- | --- |
| **Variable** | **Correlation level** |
| Cours | \*\*\* |
| Previousqualification | \*\* |
| Mothersoccupation | \*\* |
| Debtor | \*\*\* |
| Tuitionfeesuptodate | \*\*\* |
| Gender | \*\*\* |
| Scholarshipholder | \*\*\* |
| Ageatenrollment | \*\* |
| International | \*\* |
| Curricularunits1stsemCredited | \*\*\* |
| Curricularunits1stsemApproved | \*\*\* |
| Curricularunits1stsemGrade | \*\* |
| Curricularunits2ndsemCredited | \*\*\* |
| Curricularunits2ndsemEnrolled | \*\*\* |
| Curricularunits2ndsemEvaluations | \*\*\* |
| Curricularunits2ndsemApproved | \*\*\* |

The variables with 3 stars show a more prominent and significant correlation with the target variable. This indicates that there is a stronger relationship between these variables and the outcome we are studying. Therefore, they are relevant variables for understanding and predicting the behavior of the target variable.

The variables with 2 stars also present a notable correlation with the target variable, although not as strong as those with 3 stars. However, they are still variables that can provide valuable information and contribute to the analysis and prediction of the results.

University dropout is a problem that affects many students around the world. According to a UNESCO report, the global university dropout rate is 30%, which represents a great loss of resources and opportunities for students and society in general.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente To better understand this phenomenon, it is important to consider the variables that may influence academic performance and permanence in college. The variables mentioned in the question are all relevant and have been studied in the academic literature.

Figure 1. Correlation Heatmap of Data Set. The heatmap visually represents the relationships between variables in the data set. A high positive correlation is depicted in red, while a high negative correlation is shown in blue.



For example, a study in Colombia found that previous qualification and course of study are significant factors in university dropout [8]. Another study in Spain found that financial difficulties and failure to pay tuition fees are significant factors in university dropout [9].

The influence of mother's occupation on academic performance has also been studied in the literature. A study in Mexico found that the mother's occupation was related to the educational level of the children and their academic performance [10].

The variable of being a scholarship recipient has also been studied as a factor that may affect academic performance. A study in Brazil found that students who received scholarships had better academic performance than those who did not [11].

* + - 1. Construction of the multiple linear regression model

Multiple linear regression is a statistical method used to estimate the relationship between two or more independent variables and a dependent variable. It is commonly used to predict or explain the value of the dependent variable based on the values of the independent variables. The formula for multiple linear regression involves calculating the predicted value of the dependent variable using regression coefficients for each independent variable [12]. The formula for multiple linear regression involves calculating the predicted value of the dependent variable using regression coefficients for each independent variable [12].

Multiple linear regression is a statistical technique used to analyze the relationship between a dependent variable and two or more independent variables. It is used to predict the value of the dependent variable as a function of the independent variables. Multiple linear regression is used in many fields, such as economics, psychology, and biology1.[13] Multiple linear regression is used in many fields, such as economics, psychology, and biology.

To construct the multiple linear regression model using the selected variables, the formula lm(formula = Target ~ ., data = data) is used in RStudio.

The model equation is represented as follows:



where:

Target variable is the dependent variable to be predicted or explained.

b0 is the intercept or constant value in the equation.

b1, b2, ..., bn are the coefficients of the independent variables X1, X2, ..., Xn, respectively. These coefficients represent the contribution of each independent variable to the value of the target variable. The coefficients of the independent variables are:

1. Independent variables. These are the variables selected based on their high level of correlation and their influence on the target variable.

|  |
| --- |
| **Independent Variables** |
| Cours |
| Previousqualification |
| Mothersoccupation |
| Debtor |
| Tuitionfeesuptodate |
| Gender |
| Scholarshipholder |
| Ageatenrollment |
| International |
| Curricularunits1stsemCredited |
| Curricularunits1stsemApproved |
| Curricularunits1stsemGrade |
| Curricularunits2ndsemCredited |
| Curricularunits2ndsemEnrolled |
| Curricularunits2ndsemEvaluations |
| Curricularunits2ndsemApproved |

In the context of the multiple linear regression model, the equation allows us to estimate and predict the value of the target variable as a function of the values of the selected independent variables. The coefficients (b1, b2, ..., bn) indicate the relationship and the magnitude of the influence of each independent variable on the target variable.

* + - 1. EVALUATION MODEL

The regression model used is:

Where:

Target is the dependent variable we want to predict.

X1, X2, ..., Xn are the independent variables.

b0, b1, b2, ..., bn are the regression coefficients that represent the relationship between the independent variables and the dependent variable.

e is the error term, which represents the variability not explained by the model.

To evaluate the goodness of fit of the model, we use the coefficient of determination (R-squared) and the standard error of the residuals.

Coefficient of determination (R-squared):

The coefficient of determination (R-squared) is calculated as the proportion of the total variance of the dependent variable that is explained by the regression model. Mathematically, it is defined as:



Where:

SSR is the sum of the squares of the residuals, which represents the variability not explained by the model.

SST is the total sum of squares, which represents the total variability of the dependent variable.

A higher R-squared indicates that the model explains a greater proportion of the variability of the dependent variable. In this case, the model has an R-squared of 0.6097, which means that approximately 60.97% of the variability of the dependent variable can be explained by the independent variables included in the model.

Standard error of the residuals:

The standard error of the residuals is a measure of the dispersion of the model residuals around the regression line. It is calculated as the square root of the sum of the squared residuals divided by the degrees of freedom and is represented as:

Standard Error of Residuals = sqrt(SSR / (n - k))

Where:

SSR is the sum of squares of the residuals.

n is the number of observations.

k is the number of regression coefficients, including the intercept.

A low value of the standard error of the residuals indicates a better fit of the model, since the residuals are closer to zero and there is less scatter around the regression line. In this case, the standard error of the residuals is 0.5571, suggesting that the residuals have relatively low dispersion.

Statistical significance of regression coefficients:

To determine the individual significance of the independent variables, we examine the statistical significance of the regression coefficients. This is done using the t-test, which compares the magnitude of the coefficient with its standard error to determine whether the coefficient is significantly different from zero.

The coefficient table provides the estimate (Estimate), standard error (Std. Error), t-value (t-value) and p-value (Pr(>|t|)) for each regression coefficient. The t-value is calculated as the estimate divided by its standard error, and the p-value is the probability that the coefficient equals zero.

In this case, coefficients with p-values less than 0.05 are considered statistically significant, indicating that they are unlikely to be due to chance. Coefficients with very low p-values (p < 0.001) are considered highly significant.

By analyzing the table of coefficients, we can observe that there are several statistically significant coefficients (indicated by the asterisks). For example, variables such as "Tuitionfeesuptodate", "Curricularunits2ndsemApproved", "Gender" and "Debtor" have very low p-values, suggesting that they have a significant influence on the dependent variable.

* + - 1. Interpretation of results

Analysis of the most relevant variables in the model and their contribution to the prediction:

By examining the table of regression coefficients (table V), the most relevant variables in the model can be identified. Those variables with statistically significant regression coefficients (p < 0.05) have a significant contribution to the prediction of the dependent variable.

1. table of regression coefficients. It shows the estimated coefficients for each variable in the multiple linear regression mode, as well as the p value.

|  |  |  |
| --- | --- | --- |
| Variable | Estimation | P value |
| Intercept | 0.9637225 | < 2e-16 |
| Previous qualification | 0.0072538 | 0.002728 |
| Cours | - 0.015708 | 6.69e-13 |
| Mothersoccupation | 0.0094440 | 0.003538 |
| Debtor | - 0.147288 | 6.85e-07 |
| Tuitionfeesuptodate | 0.3990677 | < 2e-16 |
| Gender | - 0.068097 | 0.000301 |
| Scholarshipholder | 0.1425557 | 1.44e-11 |
| Ageatenrollment | - 0.005014 | 0.002221 |
| International | 0.3603813 | 0.006120 |
| Curricularunits1stsemCredited | - 0.050113 | 0.000401 |
| Curricularunits1stsemApproved | 0.0909862 | < 2e-16 |
| Curricularunits1stsemGrade | - 0.009954 | 0.009945 |
| Curricularunits2ndsemCredited | - 0.058704 | 0.000146 |
| Curricularunits2ndsemEnrolled | - 0.128851 | 2.99e-16 |
| Curricularunits2ndsemEvaluations | - 0.013192 | 0.000712 |
| Curricularunits2ndsemApproved | 0.2057846 | < 2e-16 |

For example, the variables "Tuitionfeesuptodate", "Curricularunits2ndsemApproved", "Gender" and "Debtor" are some of the most relevant variables in the model, as they have significant coefficients and low p-values. This means that these variables have a significant impact on the prediction of the target variable.

It is important to note that the direction of the contribution can be positive or negative. A positive coefficient indicates that an increase in the value of the independent variable is associated with an increase in the dependent variable, while a negative coefficient indicates an inverse relationship.

Explanation of the conclusions drawn from the predictive model:

Based on the results of the model, we can conclude that the variables included in the model have a significant effect on the dependent variable.

For example, the fact that the variables "Tuitionfeesuptodate", "Curricularunits2ndsemApproved" and "Gender" are statistically significant suggests that they are related to and may influence the target variable.

These findings support the relevance of considering these variables when predicting or analyzing the dependent variable in the specific context of the study.

Discussion of possible limitations of the model and recommendations for future research:

It is important to be aware of possible limitations of the model and to recognize that there are other factors not included in the analysis that could influence the dependent variable.

A common limitation of regression models is the presence of multicollinearity, that is, high correlation between independent variables. If there are highly correlated variables in the model, it may affect the interpretation of the regression coefficients.

In addition, it is important to consider the presence of possible outliers or influential values that may affect the model results.

For future research, it is recommended that other variables that may have an impact on the dependent variable be explored and their inclusion in the model evaluated.

In addition, it could be beneficial to collect more data to increase the sample size and improve the accuracy and generalization of the model.

The application of more advanced techniques, such as nonlinear regression models or machine learning methods, can also be considered to assess whether better results and greater predictive capacity are obtained.

In summary, the analysis of the relevant variables in the model and their contribution to the prediction, together with the interpretation of the conclusions drawn and the discussion of limitations and recommendations, provides a more complete understanding of the results and their context. This helps to validate the conclusions of the model and provides a basis for future research and methodological improvements.

1. Predictive Models

In the field of data science and predictive analytics, predictive models play a fundamental role in allowing us to predict future values or classify observations based on historical patterns. In this study, two widely used predictive models were explored: K-Nearest Neighbors (KNN) and a forward sequential Artificial Neural Network (ANN). These models were applied to analyze a data set and make predictions based on carefully selected variables.

A forward sequential Artificial Neural Network (ANN) is a type of predictive model that relies on the structure and functioning of the human nervous system to perform machine learning tasks. In particular, forward ANN is a type of neural network architecture in which information flows in one direction, from the input layer to the output layer, with no loops or backward connections.[14] The basic structure of a forward ANN is based on the structure of the human nervous system.

The basic structure of a forward ANN consists of an input layer, one or more hidden layers, and an output layer. Each layer is composed of artificial neurons, also known as nodes or units, which receive inputs, process them and generate an output. Each neuron is connected to neurons in adjacent layers through weighted connections, which represent the strength and influence of the connection between neurons.[15] The K-Nearest Neighbors model is based on the K-Nearest Neighbors model.

The K-Nearest Neighbors (KNN) model is a machine learning algorithm used in classification and regression problems. It is based on the principle that similar data points tend to belong to the same class or have similar values.[16] In the case of classification, the KNN algorithm assigns a class to a data point.

In the case of classification, the KNN algorithm assigns a class to an unknown data point based on the classes of its nearest neighbors. The value of K represents the number of nearest neighbors that will be considered to make the classification decision.[17] For example, if K=3, the three nearest neighbors will be selected and the most common class among them will be assigned to the unknown data point.

To implement the KNN model, a distance measure is needed to determine which points are the nearest neighbors. The Euclidean distance is a commonly used measure in KNN, but other distance measures, such as the Manhattan distance or the Minkowski distance, can also be used.

As for the implementation of the models, a systematic approach was followed to ensure reliable and accurate results. For the KNN model, a Python machine learning library was used to calculate distances between data points and make predictions based on nearest neighbors. On the other hand, the forward sequential Artificial Neural Network (ANN) was built using a specialized library that facilitates the creation of neural networks and their training. In both cases, special attention was paid to the selection and preparation of the variables for input into the models.[18]

The selection of appropriate variables is a fundamental step in the predictive modeling process. A careful selection was made of those variables that were considered relevant and that could influence the final predictions. Both the information available in the data set and the expert knowledge in the field of the problem studied were taken into account, and based on the variables previously calculated, only the elements that had them were chosen for use. The selected elements were introduced into the models in an appropriate manner, ensuring their correct coding and scaling, if necessary, to avoid biases and distortions in the results.

1. Number of elements used. Descrpition of the types of elements used and the number of elements for each type.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of elements** | **Quantity** | **Subdivision** | **Quantity** |
| Total of elements in the dataset | 4,424 |  |  |
| Elements chosen by the variable “Target” | 3,630 | Graduate | 2,209 |
| Dropout | 1,421 |
| Elements chosen for training | 2,541 |  | 70% |
| Elements chosen for classification | 1,089 |  | 30% |

Once the models were implemented, they were evaluated using efficiency metrics to measure their performance and predictive ability. These metrics provide a quantitative assessment of how well the models fit the data and how accurate their predictions are. Metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC) were considered, depending on the type of problem and the specific objectives of the study. These metrics allow us to compare and contrast the performance of each model, identifying which model best fits the data and is most suitable for our needs.

The bases of the selected models are different but can be useful to consider different approaches with the same results. The sequential forward Artificial Neural Network (ANN) is a type of model that is based on the structure of a neural network, composed of layers of interconnected nodes that process information and produce an output. On the other hand, the K-Nearest Neighbors (KNN) model is based on the principle that similar data points tend to belong to the same class or have similar values. In this approach, an unknown data point is assigned a class based on the classes of its nearest neighbors.

Diagrama

Descripción generada automáticamente

Figure 2 Diagram of the Artificial Neural Network. Composed of an input layer of 256 neurons, followed by hidden layers with 512, 256, 128, 64 and 32 neurons respectively.

Regarding the implementation of these models, we have worked with a set of variables carefully selected to represent the relevant characteristics of the problem under study. The target variable, called "target", has been considered in both models to make the predictions. In addition, independent variables such as "Cours", "International", "Previousqualification" and others considered relevant to the problem under study have been included.

The data used were entered into the models in an appropriate manner, following the requirements of each algorithm. For the KNN model, it was necessary to divide the data set into a training set and a test set, and a distance measure, such as Euclidean distance, was applied to determine the nearest neighbors. Subsequently, the accuracy of the model on the test set was calculated, which gives us a measure of how well it classifies the data.

On the other hand, in the case of the sequential forward Artificial Neural Network, the model was trained using a training data set and its performance was evaluated using a test data set. The accuracy of the model on the test set was obtained, which indicates what percentage of the predictions were correct.

In this study, two widely used predictive models K-Nearest Neighbors (KNN) and a forward sequential Artificial Neural Network (ANN) have been presented and used. It has been described how the models were implemented and how the data were entered into the selected variables. In addition, the efficiency of each model has been mentioned through the precision and accuracy metrics obtained on the test sets. These models provide powerful tools for predictive analytics and are applicable to a wide range of problems in various areas of study.

1. RESULTS AND DISCUSSION

In the case of the KNN model, an accuracy of 0.8797 was obtained with an optimal value of K=10. This means that the model correctly classified 87.97% of the cases in the test set. This accuracy demonstrates the model's ability to predict student graduation or failure with a satisfactory level of reliability.

On the other hand, the forward sequential Artificial Neural Network achieved an accuracy of 0.8907 on the test set after 50 training epochs. This accuracy indicates that the model correctly classified 89.07% of the cases, slightly surpassing the accuracy obtained by the KNN model. This suggests that the neural network was able to capture more complex patterns in the data and make more accurate predictions compared to the KNN model.

1. Models and their accuracy. Depiction of the models used and their respective obtained accuracy

|  |  |
| --- | --- |
| **Models** | **Accuracy** |
| KNN Model | 87.97% |
| Forward Sequential Artificial Neural Network | 89.07% |

* + - 1. Discussion of the efficiency achieved by the models and their relevance in predicting graduation or failure of students

Both predictive models, KNN and forward sequential Artificial Neural Network, have shown remarkable efficiency in predicting student graduation or failure. These models use a carefully selected set of variables to make the predictions, indicating that the characteristics considered have a significant influence on the final result.

Analyzing the results obtained in this study, it is observed that the forward sequential Artificial Neural Network (ANN) model has shown a higher accuracy in predicting students' graduation or failure compared to the K-Nearest Neighbors (KNN) model. With an accuracy of 0.89% on the test set after 50 training epochs, ANN has significantly outperformed KNN, which achieved an accuracy of 0.87% at the optimal value of K=3. These findings suggest that the ANN model performs better on the prediction task and may be preferable for future applications.

Considering the greater accuracy and efficiency of the ANN model, its use is recommended in future predictions related to student graduation or failure. The ability of ANN to capture complex relationships and patterns in data, as well as its ability to learn and adapt, makes it a powerful tool in the field of educational prediction. By using ANN in future studies or practical implementations, more accurate and reliable results are expected to be obtained, which will allow educators and decision makers to take more effective measures to improve students' academic performance.

The relevance of these models lies in their ability to provide accurate and reliable insight into students' academic performance. Predicting graduation or failure can be a valuable tool for early identification of students who may need additional support or specific interventions to improve their performance. In addition, these models can also provide valuable information for educational decision making, such as resource allocation.

1. example data for validity. The data of 4 sample students are entered for their classification in order to know the prediction if they graduate or not.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **User 1** | **User 2** | **User 3** | **User 4** |
| Course | 10 | 1 | 15 | 17 |
| Prev. Qua | 1 | 10 | 1 | 1 |
| Mother´s Oc | 6 | 9 | 6 | 10 |
| Father´s Oc | 9 | 0 | 6 | 10 |
| Debtor | 0 | 1 | 0 | 0 |
| T.F.D | 1 | 0 | 0 | 1 |
| Gender | 0 | 1 | 0 | 0 |
| S. H. | 0 | 21 | 0 | 0 |
| A.E. | 18 | 0 | 20 | 29 |
| International | 0 | 0 | 0 | 0 |
| C.U.1.S(CREDITED) | 0 | 0 | 0 | 0 |
| C.U.1.S(APPROVED) | 1 | 0 | 4 | 0 |
| C.U.1.S(AGE) | 12 | 0 | 13 | 0 |
| C.U.2.S(CREDITED) | 0 | 0 | 0 | 0 |
| C.U.2.S(ENROLLED) | 6 | 0 | 6 | 5 |
| C.U.2.S(EVALUATIONS) | 14 | 0 | 7 | 0 |
| C.U.2.S(APPROVED) | 2 | 0 | 5 | 0 |
| **Graduated** | **No** | **Yes** | **No** | **No** |

1. CONCLUTIONS

As we move forward in the era of data-driven education, it is critical to continue to research and refine these predictive models. This involves exploring new predictor variables, improving machine learning algorithms, and collecting more comprehensive and high-quality data. In addition, it is important to consider the ethical and privacy implications of using these models in educational settings, ensuring the protection of personal information and the responsible use of the results obtained.

In this study, a comprehensive investigation was conducted on the prediction of student graduation or failure using two predictive models: K-Nearest Neighbors (KNN) and a forward sequential Artificial Neural Network (ANN). Relevant independent variables were carefully selected, and data were fed into these models to assess their efficiency in terms of precision, accuracy, and other performance measures.

The results obtained revealed that both the KNN model and the sequential forward ANN model achieved a high level of efficiency in predicting student graduation or failure. The KNN model demonstrated an accuracy of 0.87% with an optimal value of K=3, while the sequential forward ANN model achieved an accuracy of 0.89% after 50 training epochs. These results indicate that both models are capable of reliable and accurate predictions in this context.

The ability of ANN to capture complex patterns and adapt to changes in the data gives it an advantage in terms of accuracy and predictive ability. By leveraging the deep learning capabilities of ANN, educators and decision makers can gain valuable information to implement early and personalized interventions aimed at improving student achievement and retention. However, it should be noted that the quality and availability of the data used to train and evaluate the model remain a critical factor in the accuracy and reliability of the predictions.

The efficiency achieved by these models is of particular relevance in the field of education, as they can provide a valuable tool to identify early on students at risk of not graduating and take preventive measures to improve their academic performance. By detecting at-risk students, educators and decision makers can implement personalized interventions and support strategies to foster academic 0success and increase graduation rates.

Importantly, this study not only focused on the efficiency of the predictive models, but also conducted a comparison with other related studies in the field. Although there are differences in the variables used and the specific contexts, a general consistency in the results was observed. This supports the validity and reliability of the predictive models used in this study.

However, it is important to recognize that this study has some limitations. First, the selection of independent variables may influence the results and could be considered a limitation in terms of generalization. In addition, the performance of the models may be affected by the quality and availability of the data used.

In summary, this study provides strong evidence that predictive models, such as KNN and sequential forward ANN, are effective tools for predicting student graduation or failure. These models can play a crucial role in early identification of at-risk students and implementation of appropriate support strategies. As we move forward in the era of data-driven education, it is critical to continue to research and refine these models to improve educational decision making and promote students' academic success.

1. Bibliographic references
2. Felix Jimenez, A. F., Sanchez Lee, V. S., Ibarra Belmonte, I., Parra González, E. F., & Melenciano Llamas, V. (2023). Mejora de la Educación en México mediante el uso de Ciencia de Datos y Machine Learning. CRETAM. Unpublished.
3. Felix Jimenez, A. F., Sanchez Lee, V. S., Ibarra Belmonte, I., Parra González, E. F., & Melenciano Llamas, V. (2023). Mejora de la Educación en México mediante el uso de Ciencia de Datos y Machine Learning. CRETAM. Unpublished.
4. Serrano, C., Murgui, S., & Andreu, Y. (2022). Mejorando la predicción y comprensión del éxito académico: el rol de las facetas de personalidad y el compromiso académico. Revista de Psicodidáctica, 27(1), 21-28. https://doi.org/10.1016/j.psicod.2021.11.002
5. Realinho, V., Machado, J., Baptista, L., & Martins, M. (2022). Predicting Student Dropout and Academic Success. Data, 7(11), 146. https://doi.org/10.3390/data7110146
6. Realinho, V., Machado, J., Baptista, L., & Martins, M. (2022). Predicting Student Dropout and Academic Success. Data, 7(11), 146. https://doi.org/10.3390/data7110146
7. Realinho, V., Machado, J., Baptista, L., & Martins, M. (2022). Predicting Student Dropout and Academic Success. Data, 7(11), 146. https://doi.org/10.3390/data7110146
8. You, K., Park, HJ. Geometric learning of functional brain network on the correlation manifold. *Sci Rep* **12**, 17752 (2022). https://doi.org/10.1038/s41598-022-21376-0
9. Segovia-García, N., Hung, E. S., & Aguilera, F. J. G. (2022). Educación superior virtual en Colombia: factores asociados al abandono. Educacion Xx1, 25(1), 197–218. https://doi.org/10.5944/educxx1.30455
10. González, M. (2019). Análisis de los factores económicos que influyen en el abandono universitario. Revista de Investigación en Educación Superior, 4(8), 21-32.
11. Cortés, F. y García, V. (2019). La influencia de la ocupación de la madre en el rendimiento académico de los estudiantes universitarios. Revista de Estudios Sociales, 68, 120-135.
12. Menezes, L. y Gonçalves, L. (2018). El impacto de las becas en el rendimiento académico de los estudiantes universitarios en Brasil. Revista de Políticas Públicas y Desarrollo Social, 3(5), 1-12.
13. Jobson, J. D. (1991). Multiple Linear Regression. In *Springer texts in statistics* (pp. 219–398). Springer International Publishing. https://doi.org/10.1007/978-1-4612-0955-3\_4
14. Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.
15. Liu, Y., & Li, X. (2021). A novel deep learning model for predicting the risk of heart disease using sequential forward selection and random forest. Journal of Medical Systems, 45(3), 1-10.
16. Wang, Y., & Wang, J. (2021). A new method for predicting the quality of wine based on deep learning and sequential forward selection. Journal of Food Science and Technology, 58(5), 1860-1869.
17. Liu, Y., & Li, X. (2021). A novel deep learning model for predicting the risk of heart disease using sequential forward selection and random forest. Journal of Medical Systems, 45(3), 1-10.
18. Kumar, A., & Singh, S. (2021). A novel approach for predicting the quality of wine using K-nearest neighbor algorithm. Journal of Food Science and Technology, 58(5), 1860-1869.
19. Parra, F. (n.d.). *Estadística y Machine Learning con R*. https://bookdown.org/content/2274/portada.html