Machine Learning Approach to Predict University Students' Not Completing Degree on the First Attempt based on Influential Factors

U.P. Kudagamage

2024

Machine Learning Approach to Predict University Students' Not Completing Degree on the First Attempt based on Influential Factors

A dissertation submitted for the Degree of Master of Computer Science

U.P. Kudagamage

University of Colombo School of Computing

2024

Declaration

Name of the student: U. P. Kudagamage
Registration number: 2018/MCS/048
Name of the Degree Programme: Master of Computer Science
Project/Thesis title: Machine Learning Approach to Predict University Students' Not Completing Degree on the First Attempt based on Influential Factors

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.

3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.

4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.

5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.

6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	**Date (DD/MM/YYYY)**
Upeksha	29.09.2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor 1	**Supervisor 2**	**Supervisor 3**
Name	Dr. Kasun Karunanayake		
Signature			
Date	30/09/2024		

I would like to dedicate this thesis to my parents.

ACKNOWLEDGEMENTS

First, I would like to convey my gratitude to my supervisor, Dr. Kasun Karunanayake, Senior Lecturer in the University of Colombo School of Computing for his assistance, guidance, continuous encouragement and concern over each step during this research over each step during the period of this research. Finally, I give my sincere gratitude for my loving parents, the encouragement and support they gave me during my research as well as all my colleagues in the University.

ABSTRACT

There are various types of factors that have an influence on the university students' not completing the degree on the first attempt such as financial, health or stress, academic/institutional, social and personal, economical, and disposition factors. This study's goal is to analyze the university students' decisions to complete the degree on the first attempt or not and to introduce model-based approach to predict the university students' not completing the degree on the first attempt in terms of the identified most influential factors, which will be useful in implementation of more effective individual, group-specific or institutional prevention measures. Machine learning is used for the analysis, since it has shown tremendous potential toward interpretation of complex data sets. Five different models have been trained and the trained models provided a comparatively better performance in predicting the University students' not completion of the degree on the first attempt in terms of influencing factors since all the built models gave more than 84 % of accuracy. Among them, the Naïve Bayes classifier was identified as the model with the highest of 92.75 %. An Ensemble approach was introduced and this model demonstrated an accuracy of 93.65 % which provided the best performance in predicting the University students' completion of the degree on the first attempt in terms of influencing factors considered. Further correlation coefficients which are between $r = 0.03$ and $r = 0.7$ *and* ß- coefficients which are between $r = 0.03$ and $r = 0.72$ were calculated among all the variables to determine the contribution of each variable towards the University students' not completion of the degree on the first attempt.

Key word: Not-Completing Degree, University Students, Machine Learning, Correlation Coefficients

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

University desertion and not completing the degree are considered as major problems that arises in different kinds of universities and higher educational institutions. Different controversaries in the education context were generated by this topic (Peralta, 2008). In recent days, the desertion rate is being analyzed as quality criteria university evaluation process and accreditation. Several of these cases imply academic and social changes (A. Djulovic, D. Li, 2013). The students' increase in higher education leading to a more heterogeneous student body, complicates the identification of the attributes that influence students′ decisions to complete the degree or not.

1.1. Motivation

The phenomenon of university students' not completing the degree on first attempt and the influential factors for this phenomenon are very important topics to be discussed. This leads to the inefficient use of resources and might leads to dissatisfaction of the students due to the inability to complete and achieve their educational goals. Higher educational institutions are in search of promising measures and programmes in identifying and supporting students at risk in order to minimize the wastage of human, financial and other resources. The concern of how to predict whether a particular university student is at risk of not completing the degree on the first attempt is seen around the world. Due to the introduction of new information and communication technologies, new factors have immerged that have an influence on the university students' completion of the degree on the first attempt.

Inter-related factor identification is very helpful in implementation of more effective and efficient individual or group-specific prevention measures as the ultimate factor or factors which cause the university students to decide whether to complete the degree or not on the first attempt might be unique to individuals.

1.2. Statement of the problem

When analyzing the predictive factors on the university students' completion of the degree on the first attempt, different aspects should be considered, so that it is not that easy to do the analysis. The difficulty in identifying the factors that have an influence towards the cause of university students for not completing the degree on their first attempt was demonstrated in many different researches conducted. No proper approach is demonstrated to identify the correlation of the students' not completing the degree on their first attempt to students dropping out or not completing the degree at all. Identification of the importance and the correlation of these indicators with the completion of the degree on the first attempt might be useful in many aspects such as introduction of effective measures to increase the rate of degree completion. Another problem identified is that there is no proper approach to predict the likelihood of a student not completing the degree on the first attempt in terms of the identified factors, so that more effective and efficient institutional, group-specific, or individual prevention measures for university students for not completing degree on the first attempt can be implemented.

1.3. Research Aims and Objectives

1.1.1. Aim

This study aims in identifying the most influential factors causing the university students for not completing degree on their first attempt such as financial, health or stress, academic/institutional, social and personal, economical, and disposition factors. Different types of indicators are considered in order to discover the most influential factors on the completion of the degree. Here, this study uses Machine Learning techniques for a thorough analysis of statistical relationships between influential factors and the university students' not completing the degree on their first attempt.

1.1.2. Objectives

Identifying the most influential factors causing the university students for not completing degree on their first attempt being the main objective of this research, sub-objectives are being mentioned accordingly to achieve the main objective of finding out the predictive factors that have an influence towards the cause of university students for not completing the degree on their first attempt. After discovering theses predictive factors, the correlation of students not completing the degree on the first attempt to students dropping out or not completing the degree at all with the identified different factors is to be identified. Further this study focuses on introducing a model-based approach to predict the likelihood of a student not completing the degree on the first attempt in terms of the identified factors by training different Machine Learning models and further tries to find out the predictive performance of these trained models. So that this study would implement more effective and efficient institutional, group-specific, or individual prevention measures for university students not completing degree on the first attempt.

1.4. Scope

University students' not completing the degree on the first attempt is a huge issue that should be considered. Finding out various types of factors that have an influence on the university students' not completing the degree on the first attempt and predicting the university students' not completing the degree on the first attempt in terms of the identified most influential factors with the use of Machine Learning.

1.5. Structure of the Thesis

This study evaluates various indicators or factors that have an impact on the university students' completion of the degree on the first attempt. Thesis is structured as follows. First, the introduction emphasizes the motivation for this study, the problems identified to carry out the study further and the aims and the objectives of the overall study which are to be achieved in the research. Second, the literature review highlights the various existing researches and studies carried out to in this domain as well as in similar domains with the use of different approaches such as data mining approach, statistical approach, behavioral science approach and psychological approach emphasizing the theoretical considerations and the determinants of university students not completing the degree. Further this part raises the research gap highlighting the limitations of the existing researches conducted and the contribution of this study towards the domain. Third, the methodology focuses on demonstrating the process of this research to fulfil the identified objectives. Here, the data collection and data pre-processing process, the methodologies used to train the data to predict the likelihood of a student not completing the degree on the first attempt in terms of the identified factors. Fourth, the evaluation and results illustrate the evaluation process of the trained models to determine model with the most predictive performance and the results obtained such as the most influential factors towards the university students' not completion of the degree on the first attempt and different prediction models to predict the likelihood of a student not completing the degree on the first attempt. Fifth, the conclusion and future work concludes the thesis highlighting the major findings and the conclusions of the study and further indicates the future work that can be conducted to further enhance the research.

CHAPTER 2

LITERATURE REVIEW

2.1. A Literature Review

Various existing researches and studies have been carried out to in this domain as well as in similar domains with the use of different approaches such as data mining approach, statistical approach, behavioral science approach and psychological approach emphasizing the theoretical considerations and the determinants of university students not completing the degree.

2.1.1. Theoretical considerations

A comprehensive theoretical framework is provided by two theories which explain why the university students leave a course without completing. First theory is Tinto's Student Integration Model which hypothesizes persistence to be related to how effective the individual's motivation and academic ability align with the institution's academic and social characteristics. This theory demonstrates the matching of a person's commitment to the completion of college and commitment to the institution (Tinto, 2010). Second theory is Bean's Model of Student Departure which persistence is predicted based on behavioral intentions shaped by beliefs and attitudes. Beliefs, attitudes and decisions are affected by internal factors such as students' experiences and external factors to the institution (Bean, n.d.).

2.1.2. Determinants of Students Not Completing University Education, Drop Out

One study developed personalized models for different university degrees covering Software Engineering, Humanities, Economic Sciences to obtain the risk of each student abandoning his degree and analyzes the profile for undergraduates that abandon the degree (F. Araque, C. Roldán, A. Salguero, 2009). The developed models and the framework data showed that certain variables appeared repeatedly in the explanation of the drop out in all of the faculties. These variables were start age, the father's and mother's studies, academic performance, success, average mark in the degree and the access form and in some cases also, the number of rounds needed to pass.

Another study was conducted using Decision Tree focusing on first-year university students from Portugal (J.R. Casanova, A. Cervero, J.C. Núñez, L.S. Almeida, and A. Bernardo, 2018). The results confirmed that academic performance is a determining variable in the decision making to remain or drop out, establishing three groups of achievement; high, medium and low. Here, different types of variables such as sex, type of course, the fact of being the student's first-choice university or parent's educational level. The study concluded that the weight of academic achievement should be considered as a priority variable and the identified secondary variables should be considered in the student group configuration when planning support policies to prevent higher risk of student drop-out.

The academic achievement is indicated as the determinant of students' decisions to remain on their original university degree courses in several reseraches (R. Cerezo, A.B. Bernardo, G.M. Esteban, M. Sánchez, & E. Tuero, 2015) (X. Oriol, M. Mendonza, C.G. Covarrubias, & V.M. Molina, 2017). Some researches demonstrated in their early evaluations that low academic achievement as a source of stress and dissatisfaction which increases students' disconnection from their institutions, university degree courses, and classmates (F. Belloc, a. Maruotti, & l. Petrella, 2011) (J. Gairín, X.M. Triado, M. Feixas, P. Figuera, P. Aparicio-Chueca, & M. Torrado, 2014) (K. Kinser, & J. Deitchman, 2007).

Some studies demonstrated that the prior academic histories of students were also important and showed that the students who were not entering higher education after the completion of secondary education are more likely to fail or drop out as are those students with a school history marked by situations of risk of repeating years or obtaining lower grades (Tinto, 2010) (F.M.F. Páramo, A. Araújo, C.T. Vacas, L.S. Almeida, & M.S. González, 2017). This study also demonstrated that the students with higher grades are more likely to complete their courses, especially if they enter their first-choice degree or a socially prestigious university as their previously acquired knowledge and academic competences constituting a protective factor against failure and dropout (Diseth, 2011).

The relationship of personal variables such as sex or age towards the dropout or permanence were studied. Some researchers showed that the male students spend less time on academic activities, which leads to increase their dropout rate, whereas female students who drop out tend to exhibit more difficulties with social integration. Further, they stated that the male drop-out tend to be older. This showed that for the female drop-out, age did not appear to be a determinant (P. Rosário etal., 2014), and better academic skills were demonstrated by female and it was showed that they value higher education more than male, which contrasts in the study with their peak of dropout in the case of negative results, and which seems to be more closely related to issues like balancing family and academic activities, or difficulties adapting to different assessment methods (Aina, 2013).

Since the relationship of socio-economic variables towards drop outs is more frequent in students with more disadvantaged socio-cultural background, it was also considered in some studies. These studies showed that the students whose parents had lower educational qualifications were more likely to drop out, mainly when they were first generation students, and therefore coming from families without a tradition of studying in higher education. sSome studies considered the impact of the mother's educational attainment as a factor which might have a greater impact as mother is often more present in a child's academic life and cognitive development (A. Alves, 2016) (A. Hernández etal., 2017). Further the study demonstrated that the students with more disadvantaged socio-economic backgrounds might have poor skills, poor study habits and a lack of critical thinking, which could affect their motivation and academic achievement negatively, increasing the risk of dropout (Aina, 2013) (R. Stinebrickner, & T. Stinebrickner, 2014)

Although there are many reasons why students drop out or not complete university education, those reasons may be unique for students who are enrolled in an online program. In reference (P.A. Willging & S.D. Johnson, 2020), online survey was developed to collect data from students who dropped out of an online program. Logistic regression analysis was used to compare various factors between those who persist in the program and those who dropout. Based on the dropouts from three cohorts in an online graduate program, the study showed that demographic variables did not predict likelihood of dropping from a program. Instead, the students' reasons for dropping out of an online program were varied and unique to each individual.

In reference (I. Diaz etal. , 2020), they studied dropout and transfer paths using machine learning, obtaining several key factors that were predictive for analyzing drop out and transfer paths patterns. Results showed that Polynomial SVM was the method that obtained the highest performance for predicting university dropout. Further they identified the key factors affecting university dropout, showing in addition different factors depending on the field of study.

In (M.S.A. Taipe & D.M. Sánchez, 2018), they proposed an approach to machine learning based on logistic regression techniques and decision trees and factors such as Internet addiction, addiction to social networks and addiction to technology, that affect the desertion of students in universities. As a result, it was obtained that the technique with the highest percentage of dropout precision was decision trees with 91.70%.

The reference (H. Karalar etal. , 2021) took both the synchronous and asynchronous activity characteristics of students into account to identify students at risk of academic failure during the pandemic predicted students at risk using machine learning algorithms. Performances of over two thousand university students were predicted in terms of gender, degree, number of downloaded lecture notes and course materials, total time spent in online sessions, number of attendances, and quiz score. Asynchronous learning activities were found more determinant than synchronous ones.

2.1.3. Research Gap

Many of the researches conducted were focusing on the university dropout rates, but focusing on the university students' completion of the degree on the first attempt is also important, as this would be beneficial in discovering the predictive indicators and to introduce preventive measures. Each program is unique and the reasons given for leaving a program may be specific to the nature and uniqueness of the program. This should be taken into consideration when generalizing the study.

Further, the reasons given by the students for leaving the program may be masked, due to personal issues, by an attempt to place the burden of their leaving on external factors beyond their control. Analysis of the perceptions and experiences of the instructors regarding the reasons their students left the program could help to create a more complete description of the dropout phenomenon. It would also be beneficial to include a survey of the persisters' reasons for staying in the program. The contrast between dropouts and persisters could provide further insight into the problem of university students' not completing the degree on the first attempt. Further many studies were not focusing on every aspect of factors that might be an influential factor on the not completion of the degree. And, most of the researches focuses on the students in the same discipline, though the students' discipline can also be considered as one indicator.

This study focuses on finding out various types of factors that have an influence on the university students' not completing the degree on the first attempt such as financial, health or stress, academic/institutional, social and personal, economical, disposition factors. Further this examines university students' decisions to remain or drop out of their studies or not completing on the first attempt and introduces model-based approach to predict the university students' not completing the degree on the first attempt in terms of the identified most influential factors with the use of Machine Learning by training different models, which is helpful for the implementation of more effective individual, group-specific or institutional prevention measures. Further correlation coefficients were calculated among all the variables to determine the contribution of each variable towards the University students' not completion of the degree on the first attempt.

CHAPTER 3

PROPOSED METHODOLOGY

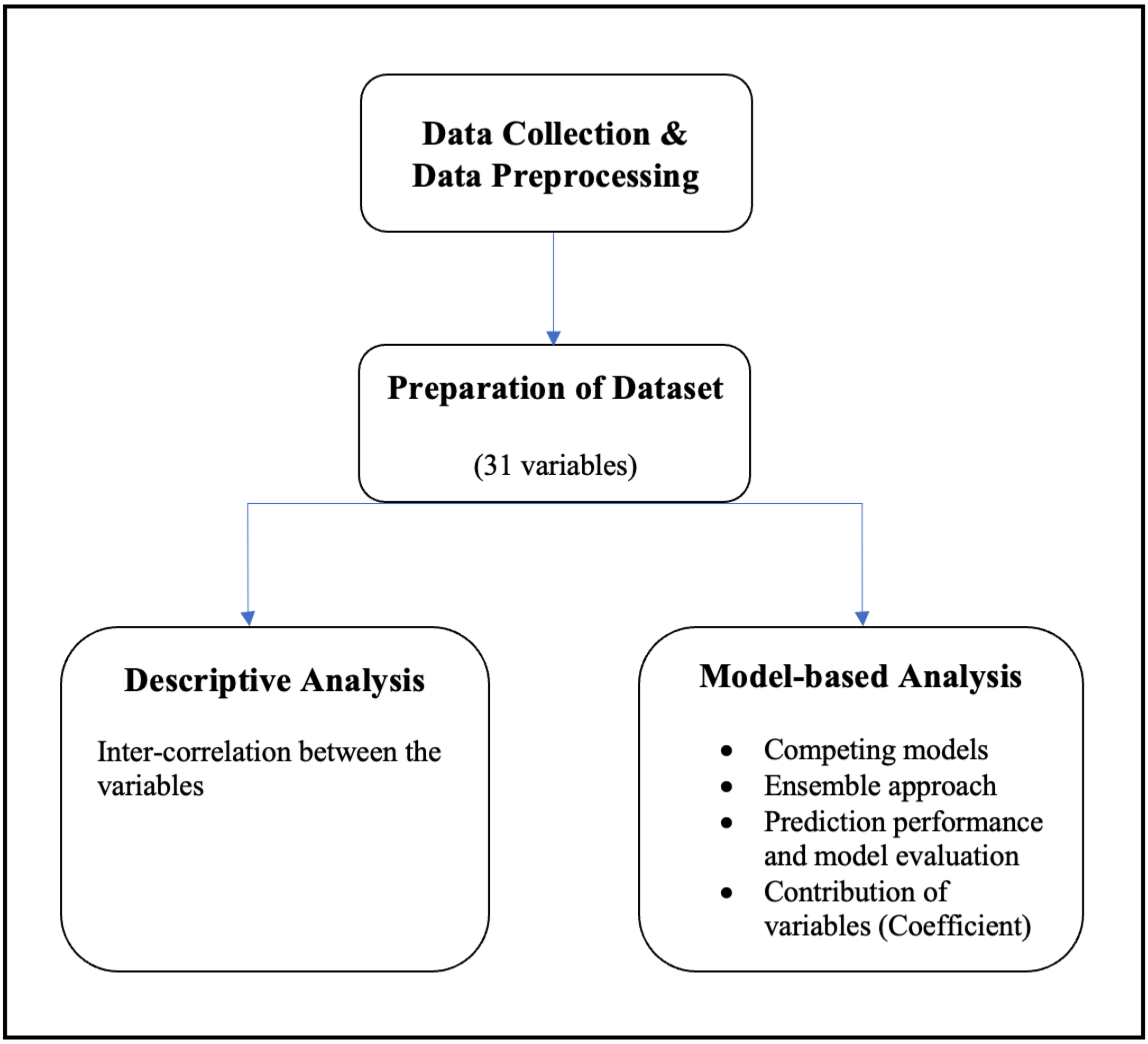

Figure 1 Flow chart of proposed methodology

3.1. Data Collection and Data Preprocessing

3.1.1 Data Collection

A questionnaire and interviews were prepared to collect the data covering the different types of factors that have an influence on the university students' not completion of the degree on the first attempt such as Academic/Institutional Factors, Psychological Factors, Social and Personal Factors, Economic Factors, Disposition (Attitude towards study) (Appendix I). This is distributed among the university graduates from different universities and from different disciplines in Sri Lanka who have completed the degree and also have not completed the degree on the first attempt. A consent form is collected from each respondent after notifying that the collected data would be kept confidential. For this analysis, six hundred and fifty (650) graduates were considered who have graduated between the period of 2012 up to 2022.

3.1.2 Data Preprocessing

In this process, handling missing values, scaling and normalization of data are done. The missing values arise due to various factors not in our direct control are handled using techniques such as imputation which replaces or fills the missing data with some value. Scaling and normalization of data is done for further analysis in which scaling changes the range of the data while normalization changes the shape of the distribution of the data.

3.2. Preparation of Dataset

This study analyzed six hundred and fifty (650) graduates collected from 2016 to 2023. Since there were some entirely irrelevant, insignificant and unimportant attributes which have less or zero contribution towards predictive modeling as compared to the critical attributes or cause a number of problems. Thirty-one (31) variables were derived and identified under the categories of Academic/Institutional Factors, Psychological Factors, Social and Personal Factors, Economic Factors, Disposition (Attitude towards study) which have an influence towards the university students' not completion of the degree on the first attempt. This study identified the university students' not completion of the degree on the first attempt as the target variable.

Table 1 Different categories of influential factors towards completion of the degree on the first attempt (the number of variables in each category is given in brackets)

Category	***Variable***
Academic/Institutional Factors (08)	Duration of the degree (three/four)
	Discipline (eleven different disciplines)
	Workload - Quizzes, Assignments etc.
	Experience in academic inadequacy
	Language barriers on study
	Student teacher involvement
	Mode of delivery (Online/physical/hybrid)

	Satisfaction with Teaching methods - Face-to-face teaching, Discussions
Psychological Factors (10)	Stress and depression Motivation Anxiety Self-esteem Self-efficacy Resilience Positive attitude Self-confidence Perfectionism
Social and Personal Factors (11)	Gender (male/female) Experience on past or recent trauma, abuse (Physical, sexual, emotional), stalking during the university period Type of personality Experience on lack of attention/ distractions (TV, Internet addiction, addiction to social networks, and addiction to technology) Expectations from family, relations, university, etc. The affectivity of family issues Difficulty in social integration Experience with relationship difficulties (emotional & physical aspects of intimate relationships) Unexpected medical problems Too many extra activities Distance from home to the university University start age
Economic Factors (01)	The affectivity of financial issues
Disposition (Attitude towards study) (01)	Being the first-choice university/degree

Under Academic/institutional factors, the duration of the degree is considered as three (03) or four (04) years and the students from different disciplines were considered such as Arts, Social Sciences, Computing, Engineering, Management, Medicine, Science, Technology, Agriculture, and Geomatics etc. The satisfaction towards the student teacher involvement is also considered as one of the influential factors as the students consider the lecturers and the instructors as mentors and tend to seek guidance from them regarding academic as well as personal matters. Due to the CORONA virus pandemic, all the academic activities were transferred into an online mode, so that the mode of delivery is also considered here. Different psychological factors such as stress and depression, motivation, anxiety, self-esteem, self-efficacy, resilience, positive attitude, self-confidence, and perfectionism are also recognized in this study, since most of the students are diagnosed with having psychological issues due to various reasons. Different types of personalities are determined here under the categories of introverted and extroverted. Due to the high expectations of the family and relations, students tend to have issues in academic activities and even develop psychological problems mentioned above. Experience with relationship difficulties in both emotional and physical aspects of intimate relationships is also one of the factors that should be linked in this study. Since the students are from different parts of Sri Lanka, they might feel homesickness due to the distance from the university.

3.3. Descriptive Analysis

Here, this study uses histograms and scatter plots to illustrate the data. By estimating the severity of multicollinearity present in the data, correlation coefficients were calculated among all the variables.

3.4. Model-based Analysis

3.4.1 Competing Models

Five Machine Learning techniques such as NaiveBayes, Decision Tree, Random Forest, Logistic Regression, and Multilayer Perceptron were trained and the prediction performance of these five models were evaluated to find out the best model to predict the likelihood of a student not completing the degree on the first attempt in terms of the identified influential factors. Theses particular models were selected as they are typically used in the existing educational researches.

NaiveBayes: The Nave Bayes algorithm is a supervised learning algorithm based on Bayes theorem which is used for solving classification problems. This Classifier is a simple and effective classification algorithm that aids in the development of fast machine learning models capable of making accurate predictions. Bayes' theorem is denoted by the formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- The probability of hypothesis A on the observed event B is P(A|B).
- Likelihood probability (P(B|A)) is the probability of the evidence given that a hypothesis' probability is true.
- Prior Probability (P(A)) is the probability of a hypothesis before it is observed.
- P(B) is Probability of Evidence Marginal Probability.

Decision Tree: Decision tree is a Supervised Learning algorithm which is used to solve regression and classification problems with the goal of creating a training model that can be used to predict the value of the target variable by learning simple decision rules inferred from prior data. It is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Random Forest: Random Forest is a supervised machine learning algorithm that is broadly used to solve classification and regression problems. It constructs decision trees from various samples and uses the majority vote for classification and the average for regression. One of the really important features of the Random Forest Algorithm is it can handle data sets with both continuous and categorical variables, as in regression and classification. When it comes to classification problems, it outperforms the competition.

Logistic Regression: Logistic Regression is a statistical model used for classification and predictive analysis. The probability of an event occurring based on a given dataset of independent variables is estimated by this algorithm. The dependant variable is bounded between 0 and 1, as the outcome is a probability. In Logistic Regression, the logit transformation is applied on the odds which is commonly known as the log odds. This logistic function is represented by the following formulas:

$$Logi(pi) = 1/(1 + \exp(-pi))$$

$$\ln\left(\frac{pi}{1 - pi}\right) = Beta_0 + Beta_1 * X_1 + \cdots + B_k * K_k$$

In this logistic equation,

- Logit(pi) = Dependant variable
- X = Independent variable
- $Beta_0$ and $Beta_1$ = The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE).

Multilayer Perceptron: A multilayer perceptron is a neural network that connects multiple layers in a directed graph, that implies that the signal only goes one way through the nodes. Aside from the input nodes, each node has a nonlinear activation function. Backpropagation is a supervised learning technique used by an MLP. MLP is a deep learning technique because it uses multiple layers of neurons. MLP is widely used in supervised learning problems, and also in computational neuroscience and parallel distributed processing research.

3.4.2 Ensemble Approach

Further this study applies an ensemble approach with the combination of the techniques that showed the best prediction performance and introduces a new model to do the predictions and increase the prediction performance and accuracy for predicting the likelihood of a student not completing the degree on the first attempt.

3.4.3 Prediction Performance and Model Evaluation

Prediction performance was assessed for these five individual models and the ensemble model by employing a 10-fold cross-validation scheme. Here, the data were randomly shuffled and divided into ten (10) data samples. The dataset was divided into test dataset and the training dataset. Here, one-tenth of the data was reserved for the testing and the remaining nine-tenth of data was used for the training of the models. The dependent variable of the reserved data was predicted by these models. In order to obtain an unbiased prediction, this process was repeated ten (10) times. After that, Accuracy, Precision, F-Measure, and Mean Squared Error (SqE) were calculated which were different criteria derived from the confusion matrix to determine the predictive performance of the models developed using different Machine learning algorithms.

Table 2 Confusion Matrix

		Actual Class	
		Positive (P)	*Negative (N)*
Predicted Class	*Positive (P)*	True Positive (TP)	False Positive (FP)
	Negative (N)	False Negative (FN)	True Negative (TN)

Table 3 Terms & Condition derived from confusion matrix

Term	***Description***
True Positive (TP)	which are correctly predicted
False Positive (FP)	which are predicted as positives, but are actually negatives
True Negative (TN)	which are predicted negatives, and they are actually negatives
False Negative (FN)	which are predicted negatives, but are actually positive
Accuracy	TP + TN / (TP + TN + FP + FN)
Precision	TP / (TP + FN)
Sensitivity	TP / (TP + FP)
F1 - Score	2 * (S * P) / (S + P)
Training dataset	dataset to build classifiers
Test dataset	dataset to evaluate the performance of the trained classifiers

3.4.4 Contribution of Variables

This study aims to determine the most contributed variables towards the prediction performance of the trained models in predicting the likelihood of a student not completing the degree on the first attempt. Here, the models were trained on the entire dataset and were analyzed to locate the variables with the most importance and significance with regard to the prediction of the target variable. For this, the ß-coefficients were calculated for the variables.

CHAPTER 4

EVALUATION AND RESULTS

This study focuses on finding the factors affecting the University students' not completion of the degree on first attempt. Six hundred and fifty (650) participants were considered in this study. After the preprocessing of the data, different Machine Learning classifiers were trained to identify the prediction performances of the models and detect the most influencing factors. The results of testing the data sets with the model are shown below. By analyzing the Accuracy, Precision, Recall, F-Measure, Root Mean Squared Error of various classification algorithms, the model that gives the highest accuracy was selected as the best prediction model.

4.1. Descriptive Analysis

The inter-correlations between the variables were computed in order to determine the multicollinearity in the data and it was discovered that the computed correlation coefficients were in the range of r = 0.03 and r = 0.7. The statistical dependance demonstrates that the significance of independent variables in predicting the target variable.

4.2. Model-based Analysis

4.2.1. Prediction Performance of Models and Analysis of Models

Dataset was prepared based on the identified factor categories. This study identified the University students' not completion of the degree on the first attempt as the target variable. Thirty-one (31) variables were determined under the categories of Academic/Institutional Factors, Psychological Factors, Social and Personal Factors, Economic Factors, Disposition (Attitude towards study) which have an influence towards the University students' not completion of the degree on the first attempt. Five different Machine Learning techniques; Naïve Bayes, Decision Tree, Random Forest, Logistic Regression and Multilayer Perceptron were trained on the dataset prepared. The prediction performance of these five individual models were determined by calculating the Accuracy, Precision, F-Measure, and Mean Squared Error (SqE).

Naïve Bayes model gave overall accuracy of 92.75 %, where the Root mean squared error is 0.2417. This model provided values of 0.865 and 0.928 for Recall and F-Measure respectively. Decision Tree model gave overall accuracy of 90 %, where the Root mean squared error is

0.2990 and Precision is 0.887. This model provided values of 0.900 and 0.889 for Recall and F-Measure respectively. Random Forest model demonstrated overall Accuracy of 85 % where the Root Mean Squared Error is 0.3689 and Precision is 0.763. This model provided values of 0.850 and 0.804 for Recall and F-Measure respectively.

When trained with the first dataset, Logistic Regression model demonstrated overall Accuracy of 92.06 %, while the Root Mean Squared Error is 0.2350 and Precision is 1.000. This model provided values of 0.855 and 0.922 for Recall and F-Measure respectively. Multilayer Perceptron demonstrated overall Accuracy of 84.15 %, while the Root Mean Squared Error is 0.376 and Precision is 0.848. The model provided the values of 0.841 and 0.844 for Recall and F-Measure respectively. Summary of performance measures of the individual models are shown in Table 4.

Table 4 Performance measures of individual models and ensemble approach

Classifier	***Accuracy***	***Precision***	***Recall***	***F-Measure***	***Root mean squared error***
Naïve Bayes	92.75 %	1.000	0.865	0.928	0.2417
Decision Tree	90.00 %	0.887	0.900	0.889	0.2990
Random Forest	85.00 %	0.763	0.850	0.804	0.3689
Logistic Regression	92.06 %	1.000	0.855	0.922	0.2350
Multilayer Perceptron	84.15 %	0.848	0.841	0.844	0.3760
Ensemble Approach (Naïve Bayes, Decision Tree, Logistic Regression)	93.65 %	1.000	0.884	0.938	0.2038

The trained models were evaluated for the performance by calculating the Accuracy, Precision, F-Measure, and Mean Squared Error (SqE). From the results obtained, it is noticed that, the five trained models have given a comparatively better performance in predicting the University students' not completion of the degree on the first attempt in terms of influencing factors since all the built models gave more than 84 % of accuracy. Among them the Naïve Bayes classifier was the model with the highest accuracy of 92.75 % and the Root mean squared error is 0.2417. The Multilayer Perceptron model provided the least accuracy of 84.15 % with a Root mean squared error of 0.3760. The introduced Ensemble model which was trained with Naïve Bayes, Decision Tree, and Logistic Regression classifiers demonstrated an accuracy of 93.65 % which surpasses the performance of the separately trained models. The proposed Ensemble Approach provided a comparatively better performance in predicting the University students' not completion of the degree on the first attempt in terms of influencing factors considered.

4.2.2. Contribution of Variables

From the data analysis of the dataset used for the prediction of completion of degree on the first attempt in terms of influencing factors, the independent variables utilized for the dataset are ranked according to its impact by using the computed correlations in the prediction of the University students' not completion of degree on the first attempt as indicated in Table 5. When considered the overall results, it is found that the affectivity of self-esteem has the highest correlation of 0.7220 towards the target variable. The variables such as Experience on past or recent trauma, abuse (Physical, sexual, emotional), stalking during the university period, Experience in academic inadequacy, and the affectivity of motivation are respectively the next highest impactful predictive variables with correlations of 0.6914, 0.6601, and 0.6370. The variable Gender was identified as the east impactful variable for the prediction of the University students' not completion of degree on the first attempt. And also, the variables such as Duration of the degree and too many extra activities are determined as the next least impactful variables for the prediction of the University students' not completion of degree on the first attempt.

Table 5 Correlation of variables towards the completion of the degree on the first attempt

	Variable	*Correlation*
1	The affectivity of self-esteem	0.7220
2	Experience on past or recent trauma, abuse (Physical, sexual, emotional), stalking during the university period	0.6914
3	Experience in academic inadequacy	0.6601
4	The affectivity of motivation	0.6370
5	The affectivity of anxiety	0.5950
6	The affectivity of self-confidence	0.5480

7	Experience with relationship difficulties (emotional & physical aspects of intimate relationships)	0.5450
8	The affectivity of self-efficacy	0.5350
9	The affectivity of resilience	0.5190
10	The affectivity of stress and depression	0.5060
11	The affectivity of Perfectionism	0.5000
12	The affectivity of positive attitude	0.4650
13	Unexpected medical problems	0.4317
14	Difficulty in social integration	0.3946
15	Distance from home to the university	0.3494
16	Student teacher involvement	0.3131
18	Discipline	0.3091
19	Type of personality	0.295
20	Workload - Quizzes, Assignments	0.2862
21	The affectivity of financial issues	0.2825
22	The affectivity of family issues	0.2712
23	Satisfaction with Teaching methods - Face-to-face teaching, Discussions	0.2459
24	Expectations from family, relations, university, etc.	0.2299
25	Mode of delivery	0.2185
26	Experience on lack of attention/ distractions (TV, Internet addiction, addiction to social networks, and addiction to technology)	0.2107
27	The affectivity of Language barriers on study	0.1921
28	Start age	0.1714
29	Too many extra activities	0.1457
30	Duration of the degree	0.1115
31	Gender	0.0347

When considering the Academic/Institutional Factors, the variable Experience in academic inadequacy is determined as the highest impactful academic/institutional factor with a correlation of 0.6601 for the prediction of the University students' not completion of degree on the first attempt. And the variables Student teacher involvement and Discipline are the next most impactful academic/institutional factors with correlations of 0.3131 and 0.3091 respectively.

When the Psychological factors utilized for the dataset are ranked according to its impact or correlation in the prediction of completion of degree on the first attempt, it is found out that the variable, The affectivity of self-esteem was determined as the most influential psychological factor with a correlation of 0.7220 and it has the highest correlation when considered all the categories. The variables such as the affectivity of motivation and the affectivity of anxiety are the second and third highly impactful psychological factors with correlations of 0.6370 and 0.5950 respectively. The affectivity of positive attitude and the affectivity of Perfectionism are the least impactful psychological factors for the prediction of the University students’ not completion of degree on the first attempt with correlations of 0.4650 and 0.5000 respectively.

The Experience on past or recent trauma, abuse (Physical, sexual, emotional), stalking during the university period has the highest contribution towards the prediction of the University students’ not completion of degree on the first attempt with a correlation of 0.6914. The Experience with relationship difficulties (emotional & physical aspects of intimate relationships) is the next most impactful social and personal factor with a correlation of 0.5450. Gender is the least contributed social and personal factor with a correlation of 0.0347 and it is the least impactful factor when considered all the categories of predictive factors.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This study focuses on finding out various types of factors that have an influence on the university students' not completion of the degree on the first attempt such as financial, health or stress, academic/institutional, social and personal, economical, disposition factors. Further this examines the university students' decisions to remain or drop out of their studies or not completing on the first attempt and introduces model-based approach to predict the university students' not completion of the degree on the first attempt in terms of the identified most influential factors, which is helpful for the implementation of more effective individual, group-specific or institutional prevention measures.

This study analyzed six hundred and fifty university students who have graduated and identified thirty-one factors which have an impact on the target variable. Five different Machine Learning models have been trained and the trained models provided a comparatively better performance in predicting the university students' completion of the degree on the first attempt in terms of influencing factors, since all the built models were evaluated with more than 84 % of accuracy. First result of the study is determining the model with the best prediction performance. Among these trained models, the Naïve Bayes model was identified as the model with the highest accuracy of 92.75 % and the Multilayer Perceptron model was identified as the model with the least accuracy of 84.15 %. Further an Ensemble approach was introduced and this model was trained with Naïve Bayes, Decision Tree, and Logistic Regression classifiers which demonstrated an accuracy of 93.65 % which provided the best performance in predicting the University students' completion of the degree on the first attempt in terms of influencing factors considered. Second result of this study is that the affectivity of self-esteem was identified as most influential factor among the all types of factors considered in the study with a correlation of 0.7220 towards the University students' not completion of the degree on the first attempt. Further this study examines that the correlation of the factors identified are withing the range of $r = 0.03$ and $r = 0.72$.

As for the future work, different classifiers will be considered for building the prediction models to discover the model with the best performance. And further more statistical approach can be applied to discover the variation of the data and to analyze how the data is distributed.

APPENDICES

Appendix I

Questionnaire to identify the factors affecting the university students not completing the Degree Programme on the first attempt

Questionnaire to identify the factors affecting the university students not completing the Degree Programme on the first attempt

This survey is conducted as a partial requirement of a Master's Degree Programme with the main purpose of identifying the factors identify the factors affecting the university students not completing the Degree Programme on the first attempt. You are kindly requested to participate in this analysis and give your honest responses.

All responses to this questionnaire will be held and treated in strict confidence and used only for analysis purposes. Please provide background information about yourself. Thank you very much for taking the time to complete this questionnaire, and your contribution is very much appreciated.

* Required

Part I: General Information

1. Gender *

Mark only one oval.

- ⬭ Male
- ⬭ Female
- ⬭ Other

2. Age *

Mark only one oval.

- ⬭ Below 20
- ⬭ 21 - 24
- ⬭ 25 - 28
- ⬭ Above 29

3. Title *

Mark only one oval.

- Prof.
- Dr.
- Mr.
- Mrs.
- Ms.

4. Type of Degree *

Mark only one oval.

- Honours
- Special
- General
- Other

5. Duration of the degree *

Mark only one oval.

- Below 2 years
- 2 years
- 3 years
- 4 years
- Above 4 years

6. Discipline *

Mark only one oval.

- Arts
- Social Sciences
- Computing
- Engineering
- Management
- Medicine
- Science
- Technology
- Agriculture
- Geometrics
- Other

7. Did you complete the degree on the first attempt? *

Mark only one oval.

- Yes
- No

8. In which attempt did you complete the degree programme? *

Mark only one oval.

- 1st Attempt
- 2nd Attempt
- 3rd Attempt

9. Does the inability to complete the degree on the first attempt affect your career? *

Mark only one oval.

- Yes
- No
- Not applicable

Academic/Institutional Factors

10. Workload - Quizzes, Assignments *

Mark only one oval.

- Very low
- Low
- Average
- High
- Very high

11. Experience in academic inadequacy *

Mark only one oval.

- Yes
- No
- Neutral

12. The affectivity of Language barriers on study *

Mark only one oval.

- Very Low
- Low
- Average
- High
- Very high

13. Satisfaction with Teaching methods - Face-to-face teaching, Discussions *

Mark only one oval.

- Very unsatisfied
- Unsatisfied
- Neutral
- Satisfied
- Very Satisfied

14. Difficulty in adapting to different assessment methods *

Mark only one oval.

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

15. Student teacher involvement *

Mark only one oval.

- Very poor
- Poor
- Average
- Good
- Very good

16. Academic achievement in their early evaluations *

Mark only one oval.

- Very low
- Low
- Average
- High
- Very high

17. How often did you have to repeat subjects *

Mark only one oval.

- Zero or very rare
- Rare
- Neural
- Often
- Very often

18. Attendance for lectures and assessment *

Mark only one oval.

- Very poor
- Poor
- Average
- Good
- Very good

19. Mode of delivery *

Mark only one oval.

- Physical
- Online
- Hybrid

20. Feeling like falling behind in your studies? *

Mark only one oval.

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

21. Time management *

Mark only one oval.

- Very poor
- Poor
- Avearge
- Good
- Very good

22. Experience on lack of attention/ distractions (TV, Internet addiction, addiction to social networks, and addiction to technology)

Mark only one oval.

- Zero or very rare
- Rare
- Neutral
- Often
- Very often

23. Satisfaction with the availability of resources – Computers, books, reading materials, lecture materials

Mark only one oval.

- Very unsatisfied
- Unsatisfied
- Neutral
- Satisfied
- Very satisfied

24. Satisfaction with the physical study environment - noisy or lacking privacy *

Mark only one oval.

- Very unsatisfied
- Unsatisfied
- Neutral
- Satisfied
- Very satisfied

Psychological Factors (especially in academic activities)

25. The affectivity of stress and depression *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

26. The affectivity of motivation *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

27. The affectivity of anxiety *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

28. The affectivity of self-esteem *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

29. The affectivity of self-efficacy *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

30. The affectivity of resilience *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

31. The affectivity of positive attitude *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

32. The affectivity of self-confidence *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

33. The affectivity of Perfectionism *

Mark only one oval.

- Very low
- Low
- Neutral
- High
- Very high

Social & Personal Factors

34. Experience on past or recent trauma, abuse (Physical, sexual, emotional), stalking during the university period

Mark only one oval.

- Yes
- No
- Neutral

35. Type of personality *

Mark only one oval.

- Introverted thinking
- Extraverted thinking
- Introverted feeling
- Extraverted feeling
- Introverted sensation
- Extraverted sensation
- Introverted intuition
- Extraverted intuition

36. Expectations from family, relations, university, etc. *

Mark only one oval.

- Very low
- low
- Neutral
- High
- Very high

37. The affectivity of cultural differences *

Mark only one oval.

- Very low
- low
- Neutral
- High
- Very high

38. The affectivity of family issues *

Mark only one oval.

- Very low
- low
- Neutral
- High
- Very high

39. Conflict with study and family commitments (having to help at home reduce the time available for study)

Mark only one oval.

- Very low
- low
- Neutral
- High
- Very high

40. Father's and mother's educational background *

Mark only one oval.

- Very poor
- Poor
- Average
- Good
- Very good

41. Support from family *

Mark only one oval.

- Very poor
- Poor
- Average
- Good
- Very good

42. Homesickness and feeling that you don't fit in *

Mark only one oval.

- Yes
- No
- Neutral

43. Difficulty in social integration *

Mark only one oval.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly agree

44. Discouraging environment- Lack of student engagement *

Mark only one oval.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly agree

45. No guidance or mentors for both academic and personal matters *

Mark only one oval.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly agree

46. Experience with relationship difficulties (emotional & physical aspects of intimate relationships)

Mark only one oval.

- Yes
- No
- Neutral

47. The affectivity of life events (loss of a loved one/ relationship breakup) *

Mark only one oval.

- Yes
- No
- Neutral

48. How often did you have unexpected medical problems? *

Mark only one oval.

- No or very rare
- Rare
- Neutral
- Often
- Very often

49. Too many extra activities *

Mark only one oval.

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

50. Distance from home to the university *

Mark only one oval.

- Below 5 km
- 5 - 30 km
- 31 - 60 km
- 61 - 90 km
- 91 - 120 km
- Above 120 km

Economic Factors

51. How did you manage living costs during the university period? *

Mark only one oval.

- Self-funded by working
- Self-funded by parents/ siblings/ relatives / friends etc.
- Bursary
- Scholarship
- Other

52. How often did you face financial issues? *

Mark only one oval.

- No or very rare
- Rare
- Neutral
- Often
- Very often

53. Did you do part-time work during the university period? *

Mark only one oval.

- Yes
- No
- Neutral

Disposition (attitude towards study)

54. Start age *

Mark only one oval.

- Below 20 years
- 21 - 23 years
- 24 - 26 years
- 27 - 29 years
- Above 29 years

55. Is this your first-choice university? *

Mark only one oval.

- Yes
- No
- Neutral

56. Feeling that you had chosen the wrong major *

Mark only one oval.

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

57. Overall satisfaction with the university and degree programme *

Mark only one oval.

- Very unsatisfied
- Unsatisfied
- Neutral
- Satisfied
- Very satisfied

REFERENCES

[1] F. Araque, C. Roldán, A. Salguero, 2009. Factors influencing university drop out rates. *Computers & Education,* November.

[2] J.R. Casanova, A. Cervero, J.C. Núñez, L.S. Almeida, and A. Bernardo, 2018. Factors that determine the persistence and dropout of university students. *Psicothema,* Volume Vol. 30.

[3] R. Cerezo, A.B. Bernardo, G.M. Esteban, M. Sánchez, & E. Tuero, 2015. Programs for promoting self-regulated learning in higher education: A study of the satisfaction between in-person and virtual methods. *European Journal of Education and Psychology,* Volume 8(1), pp. 30-36.

[4] X. Oriol, M. Mendonza, C.G. Covarrubias, & V.M. Molina, 2017. Positive emotions, autonomy support and academic performance of university students: The mediating role of academic engagement and self-efficacy. *Revista de Psicodidáctica,* pp. 45-53.

[5] F. Belloc, a. Maruotti, & l. Petrella, 2011. How individual characteristics affect university students drop-out: A semi-parametric mixed-effects model for an Italian case study. *Journal of Applied Statistics,* pp. 2225-2239.

[6] J. Gairín, X.M. Triado, M. Feixas, P. Figuera, P. Aparicio-Chueca, & M. Torrado, 2014. Student dropout rates in Catalan universities: Profile and motives for disengagement. *Quality in Higher Education,* pp. 165-182.

[7] K. Kinser, & J. Deitchman, 2007. Tenacious persisters: Returning adult students in Higher Education. *Journal of College Student Retention: Research, Theory & Practice,* pp. 75-94.

[8] Tinto, V., n.d. From theory to action: Exploring the institutional conditions for student retention. *Higher Education: Handbook of Theory and Research,* Volume Vol. 25, pp. 51-89.

[9] F.M.F. Páramo, A. Araújo, C.T. Vacas, L.S. Almeida, & M.S. González, 2017. Predictors of students' adjustment during transition to university in Spain. *Psicothema,* pp. 67-72.

[10] P. Rosário etal., 2014. An explanatory model of the intention to continue studying among non- traditional university students. *Psicothema,* pp. 84-90.

[11] Aina, C., 2013. Parental background and university dropout in Italy. Higher Education. pp. 437-456.

[12] A. Alves, 2016. Cognitive performance in childhood: The mother and the urban environment make a difference]. *Psicologia: Teoria e Pesquisa,* pp. 1-9.

[13] A. Hernández etal., 2017. he effect of demographic variables on the assessment of cognitive ability. *Psicothema,* pp. 469-474.

[14] R. Stinebrickner, & T. Stinebrickner, 2014. A major in Science? Initial beliefs and final outcomes for college major and dropout. In: *Review of Economic Studies.* s.l.:s.n., pp. 426-472.

[15] P.A. Willging & S.D. Johnson, 2020. FACTORS THAT INFLUENCE STUDENTS' DECISION TO DROPOUT OF ONLINE COURSES. *Journal of Asynchronous Learning Networks,* 13(3).

[16] Tinto, V., 2010. From theory to action: Exploring the institutional conditions for student retention. *Higher Education: Handbook of Theory and Research,* Volume Vol. 25, pp. 51-89.

[17] Bean, J., n.d. Why students leave: Insights from research. In: *The Strategic Management of College Enrollments.* s.l.:s.n.

[18] I. Diaz etal. , 2020. *Variables Influencing University Dropout: A Machine Learning-Based Study.* s.l., s.n., pp. 94-103.

[19] Peralta, D., 2008. Conceptual model for dropout chilean university student. *Estudios pedagógicos,* pp. 65-86.

[20] A. Djulovic, D. Li, 2013. Towards freshman retention prediction: a comparative study. *International Journal of Information and Education Technology,* p. 494.

[21] M.S.A. Taipe & D.M. Sánchez, 2018. Prediction of university dropout through technological factors: a case study in Ecuador.

[22] H. Karalar etal. , 2021. Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education.*

[23] Diseth, A., 2011. *Self-efficacy, goal orientations and learning strategies as mediators between preceding and subsequent academic achievement.* s.l., s.n., pp. 191-195.