Fundamentals of Data Science & Basic Statistics Final Project Report

Lecturer:

Nunung Nurul Qomariyah, S.Kom., M.T.I., Ph.D.

Dr. Raymond Bahana, ST., M.Sc

Project Name:

Predicting Student Dropout and Factors that Correlate to Dropping Out

Group Members:

Edward Raphael - 2702355391

Harry Chiu - 2702357882

Ahmad Zaydan - 2702358393

Class: L3CC

Fundamentals of Data Science (COMP6784001)

Basic Statistics (STAT6171001)

FACULTY OF COMPUTER SCIENCE

BINUS INTERNATIONAL UNIVERSITY

2024

Table of Contents

I. Abstract

In the pursuit of advancement in society, the youth must be able to adapt and transition into the real world. The world within the workforce is a rough and unforgiving space, thus the new generation must be equipped with the right mentality and skills to survive. With the new trend of dropping out of universities, students will lack the skills required to thrive. Our research can counter against this wave by utilizing a dataset that includes student attributes and applying 5 machine learning techniques to create predictive models for universities. Such techniques like random forest, XGBoost, decision tree, support vector machine and k-nearest neighbor. Our models are designed to predict whether or not a student will drop out, and determine the factors that contribute to a drop out. The results show that tuned XGBoost is the model with the highest balanced accuracy of 0.921. Along with that, we also combined the models into a VotingClassifier and it produced comparable accuracies to the XGBoost with a value of 0.922.

II. Problem Analysis

In this modern era, the topic of dropping out is no uncommon matter. Students are dropping out more than ever and it is a concern. High dropout rates can affect the community in many ways. Research states that a single high school dropout can cost the US economy $250,000 throughout their whole life [1]. The reasoning to that is the probable decrease in earnings of civilians, increased reliance on government welfare and also the increase in criminal rates.

By addressing the dropout rates, we can benefit our society in many ways. The first one being the improvement in financial earnings. Students can enhance their future earnings by remaining in university to develop their skills, also to contribute to the workforce [2]. Secondly, it reduces the risk of being health deprived. Being educated often leads to better and healthier life choices [1]. Thirdly, in the long run, not dropping out reduces rates of poverty. Thus, enhances the economic status of the community for future generations to come.

By formulating models to predict the conditions of the students, we can comfortably assure that our society is going in a positive path. The aim of the model is to predict which attributes contribute the most to dropping out and prevent any incidents by predicting the outcomes beforehand through the student's features and attributes.

III. Related Work (Literature Review)

A. Factors Affecting Student Drop-Out Behavior

Student drop-out rates pose significant challenges for educators and policymakers globally. This literature review prioritizes recent empirical studies from 2015 to 2023 to identify key factors contributing to student drop-out behavior and evaluating measures attempting to mitigate these rates. The research highlights the key factors influencing student drop-outs, including family related factors, the research constantly states the importance of parental involvement in education with studies indicating that active participation of parents correlates to lower drop out rates or the socioeconomic status of the family where limited access to resources likely increase the rate of dropouts. Negative school experience

diminishes students' motivation which leads to poor academic performance, another strong indicator of dropout behavior. Other factors include social dynamics which could significantly impact students where negative peer pressure could lead to students being discouraged to pursue higher education. The paper suggests some solutions to intervene dropout rates, tailored support programs which are remedial programs that address specific problems faced by the students. Another solution is by enhancing school support services by providing students with comprehensive guidance and counseling programs that can provide students with necessary resources to navigate through academic challenges and personal issues.[3]

B. **Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods†**

Various studies have been done before to predict student dropouts such as work by Bowers, Sprott, and Taff (2013). The studies done on student's data to predict possible students at risk reveals important findings and key studies which are essential in determining the factors that influence a student's drop out rate. One of the key insights to predict a student's attrition is their early performance data, with the availability of their data the demographic information has limited predictive value and thus factors such as age, gender, ethnicity, and socioeconomic status do not significantly enhance the ability to predict whether a student will drop out of their educational program. This paper also discusses the theoretical framework, where Bean's (1983) turnover model, originally developed for organizations, has been adapted to analyze student attrition in higher education, offering a theoretical framework for understanding the reasons behind students' decisions to drop out of higher education. There has also been usage of data mining in this study, Zhang et al. (2010) showcases this by analyzing historical data, educational institutions can uncover trends and factors that contribute to student attrition, allowing institutes to make data-driven choices that can positively impact the students.[4]

C. **Global Challenges of Students Dropout: A Prediction Model Development Using Machine Learning Algorithms on Higher Education Dataset**

Due to the rapid development in data availability in the world, the ability to analyze patterns from a diversity of datasets have expanded. By taking advantage of this understanding, Meseret Yihun Amare and Stanislava Simonova (2021) from the University of Pardubice utilizes machine learning methods to identify cases of early dropouts for University students. By addressing this issue, institutions may improve retention rates and the general growth of educational systems. The study discovered logistic regression as the model with the highest performance in terms of accuracy, precision, recall and F1-Score. Out of the 90 test records they used for training their models, the model managed to identify 71 non-dropouts and 14 real dropouts. The average rate for error for their model is merely 0.056%. [5]

D. Factors that determine the persistence and dropout of university students

This paper focuses on the impact of diverse socio-demographic backgrounds, academic achievement, and institutional characteristics on retention rates. The access to higher education resulted in a varied student population where academic performance, particularly grade point averages (GPAs) and first-year achievements, prove to be a primary factor of retention. Socio-demographic factors such as age, gender and parental teachings significantly influence the students' academic goals. Moreover, the choice of field the students study in has an effect with dropout rates, with integrated master's programs often demonstrating different retention patterns compared to shorter degrees. Paper also mentioned the importance of self-regulation and engagement which are vital to achieve academic success. Overall, it highlights the complexity of these diverse influences and for prevention to be made, it will require interventions that will consider the different backgrounds and factors to advocate for a better and inclusive learning environment to drop retention rates. [6]

E. Understanding the Complex Factors behind Students Dropping Out of School

This paper explores the multifaceted reasons contributing to student dropout rates, emphasizing that this issue is not static but a dynamic process influenced by various factors over time. Research indicates that students' engagement with

academic curriculum and social dynamics begin when they enter an educational system. Many students expressed the separation between their learning materials and real world application, leading some of them to consider other paths in life. This highlights the need for education to be meaningful and relevant. Parental influence is a major factor, studies show that parental belief and decisions significantly affect the students' likelihood of dropping out. A supportive family is essential to provide more positive experiences. Additionally, school environments also play a major role in shaping a student, a supportive environment would foster and enhance the students' motivation and commitment. This study promotes connecting classroom education with real-world applications to help students appreciate the value of education. In conclusion, the complex relationship between individual motivations, family influence and social environment significantly affect the students' choice to stay in or out of school. By engaging a comprehensive approach, educators and policymakers can create effective strategies to tackle dropout rates. [7]

IV. Data Set and Preprocessing

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4424 entries, 0 to 4423
Data columns (total 35 columns):

#	Column	Non-Null Count	Dtype
0	Marital status	4424 non-null	int64
1	Application mode	4424 non-null	int64
2	Application order	4424 non-null	int64
3	Course	4424 non-null	int64
4	Daytime/evening attendance	4424 non-null	int64
5	Previous qualification	4424 non-null	int64
6	Nacionality	4424 non-null	int64
7	Mother's qualification	4424 non-null	int64

```
 8   Father's qualification                        4424 non-null   int64
 9   Mother's occupation                           4424 non-null   int64
 10  Father's occupation                           4424 non-null   int64
 11  Displaced                                     4424 non-null   int64
 12  Educational special needs                     4424 non-null   int64
 13  Debtor                                        4424 non-null   int64
 14  Tuition fees up to date                       4424 non-null   int64
 15  Gender                                        4424 non-null   int64
 16  Scholarship holder                            4424 non-null   int64
 17  Age at enrollment                             4424 non-null   int64
 18  International                                 4424 non-null   int64
 19  Curricular units 1st sem (credited)           4424 non-null   int64
...
 33  GDP                                           4424 non-null   float64
 34  Target                                        4424 non-null   object
dtypes: float64(5), int64(29), object(1)
memory usage: 1.2+ MB
```

Data Preprocessing

```
students.rename(columns = {"Nacionality": "Nationality",
                    "Mother's qualification":"Mother_qualification",
                     "Father's qualification""Father_qualification",
                     "Mother's occupation": "Mother_occupation",
                     "Father's occupation": "Father_occupation",
                     "Age at enrollment": "Age"}, inplace = True)

students.columns = students.columns.str.replace(' ', '_')
students.columns = students.columns.str.replace('(', '')
students.columns = students.columns.str.replace(')', '')
```

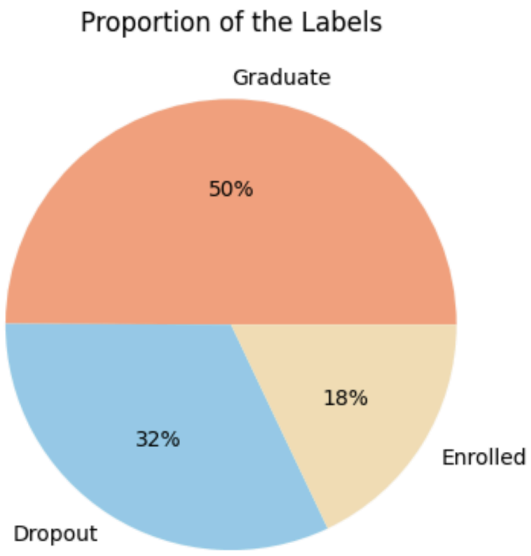
The data is mostly cleaned, but there are still some inconsistencies in the columns. First we replace all the typos and replace all spaces with underscores. Then, remove any white spaces with underscores.

```
col = ['Marital_status', 'Application_mode', 'Application_order', 'Course',
    'Daytime/evening_attendance', 'Previous_qualification', 'Nationality',
     'Mother_qualification', 'Father_qualification', 'Mother_occupation',
     'Father_occupation', 'Displaced', 'Educational_special_needs', 'Debtor',
     'Tuition_fees_up_to_date', 'Gender', 'Scholarship_holder',
    'International', 'Target']

students[col] = students[col].astype('category')
```

Change data types of columns from Int to category for the models to treat the data as categorical.

```
# Check the proportion of the labels in the target variable
labels = students['Target'].value_counts().index
values = students['Target'].value_counts().values

plt.pie(values, labels = labels, colors = ['lightsalmon', 'skyblue', 'wheat'],
        autopct = '%1.0f%%')
plt.title('Proportion of the Labels');
```

This code produces the table showing the percentage of the “Target” from the students



From this table, we can identify that the majority of university students completed their degree and graduated, however, there are still a lot of students who drop out from university.

```
# Encode the labels as ordinal data (0 - 'Dropout', 1 - 'Enrolled', and 2 - 'Graduate')
students['Target_encoded'] = OrdinalEncoder(categories = [['Dropout', 'Enrolled', 'Graduate']]).fit_transform(students[['Target']])

# Drop 'Target' variable
students.drop('Target', axis = 1, inplace = True)
```

Change the values in the “Target” variable into ordinal values. 0 for Dropout, 1 for enrolled, and 2 for graduate

```
# list of categorical features
cats = ['Marital_status', 'Application_mode', 'Application_order',
        'Course','Daytime/evening_attendance', 'Previous_qualification',
        'Nationality','Mother_qualification', 'Father_qualification',
        'Mother_occupation', 'Father_occupation', 'Displaced',
        'Educational_special_needs', 'Debtor','Tuition_fees_up_to_date',
        'Gender', 'Scholarship_holder','International']

# Get the p-values from Chi-Square indenpendence tests
p_value = []

for col in cats:
    crosstable = pd.crosstab(index = students[col],
                             columns = students['Target_encoded'])
    p = chi2_contingency(crosstable)[1]
    p_value.append(p)

chi2_result = pd.DataFrame({
    'Variable': cats,
    'P_value': [round(ele, 5) for ele in p_value]
})

chi2_result = chi2_result.sort_values('P_value')

chi2_result
```

This part of the code calculates the p-values using Chi-square independence test.

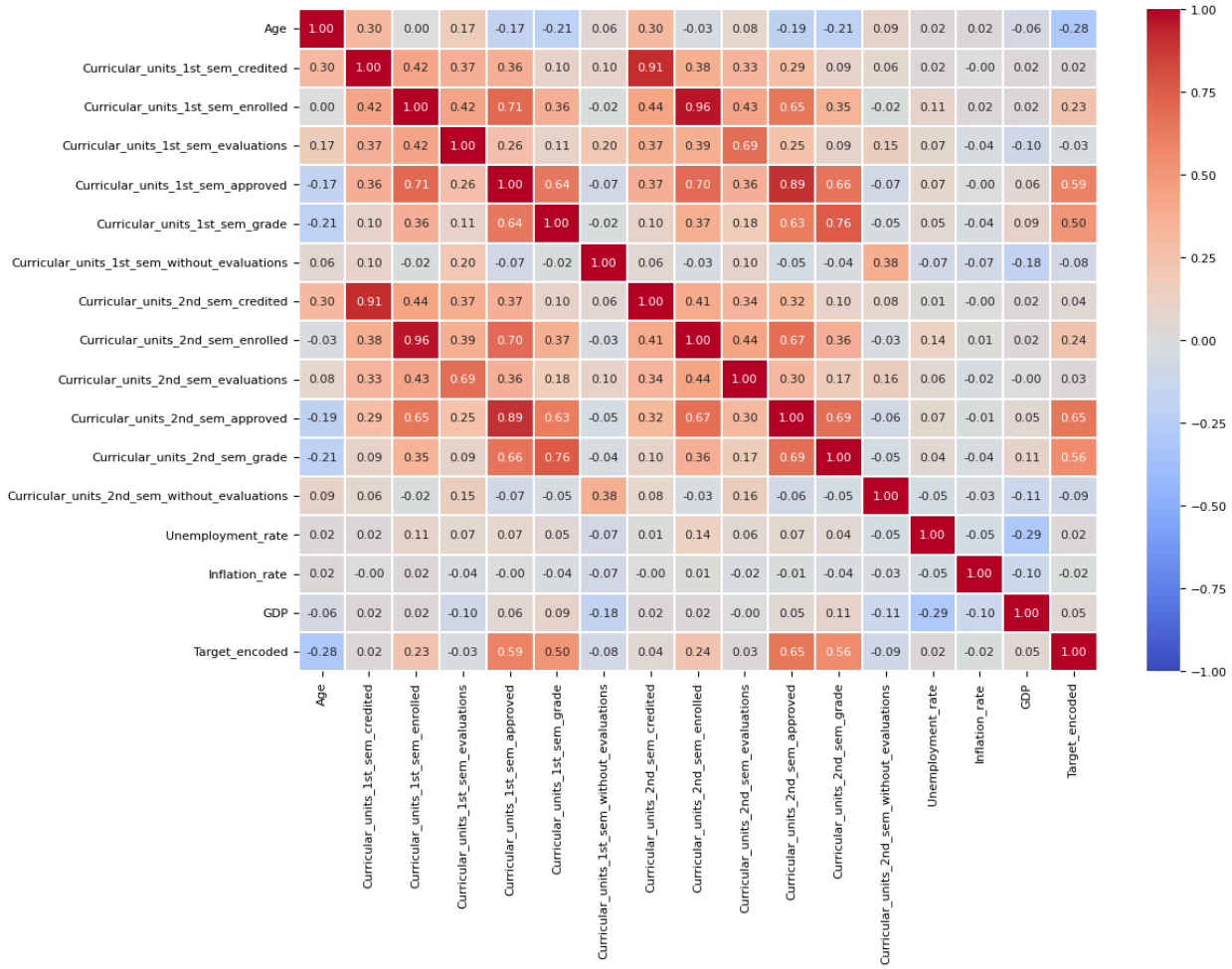
	Variable	P_value
0	Marital_status	0.00000
15	Gender	0.00000
14	Tuition_fees_up_to_date	0.00000
13	Debtor	0.00000
11	Displaced	0.00000
10	Father_occupation	0.00000
9	Mother_occupation	0.00000
16	Scholarship_holder	0.00000
8	Father_qualification	0.00000
5	Previous_qualification	0.00000
4	Daytime/evening_attendance	0.00000
3	Course	0.00000
2	Application_order	0.00000
1	Application_mode	0.00000
7	Mother_qualification	0.00000
6	Nationality	0.24223
17	International	0.52731
12	Educational_special_needs	0.72540

```
stud_selected = students.drop(['Nationality', 'International', 'Educational_special_needs'], axis = 1)
```

```
# Numeric features
num_features = students[['Age',
       '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',
       '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',
       'Unemployment_rate', 'Inflation_rate', 'GDP', 'Target_encoded']]

# Heatmap of correlation matrix
plt.figure(figsize = (12, 8))
plt.rcParams.update({'font.size': 8})
hm = sns.heatmap(num_features.corr(method = 'spearman'),
                 cmap = 'coolwarm', annot = True, fmt = '.2f',
                 linewidths = .2, vmin = -1, vmax = 1, center = 0)
```

This part of the code generates a heatmap to visualize the Spearman's rank correlation between numerical features and the labels. Spearman's rank correlation measures the strength and direction of monotonic association between two variables. It can capture both linear and nonlinear monotonic relationships.

Here is the heatmap:



As can be seen from the heat map, there are four features:

- 'Curricular_units_2nd_sem_approved',
- 'Curricular_units_2nd_sem_grade',
- 'Curricular_units_1st_sem_approved',
- 'Curricular_units_1st_sem_grade')

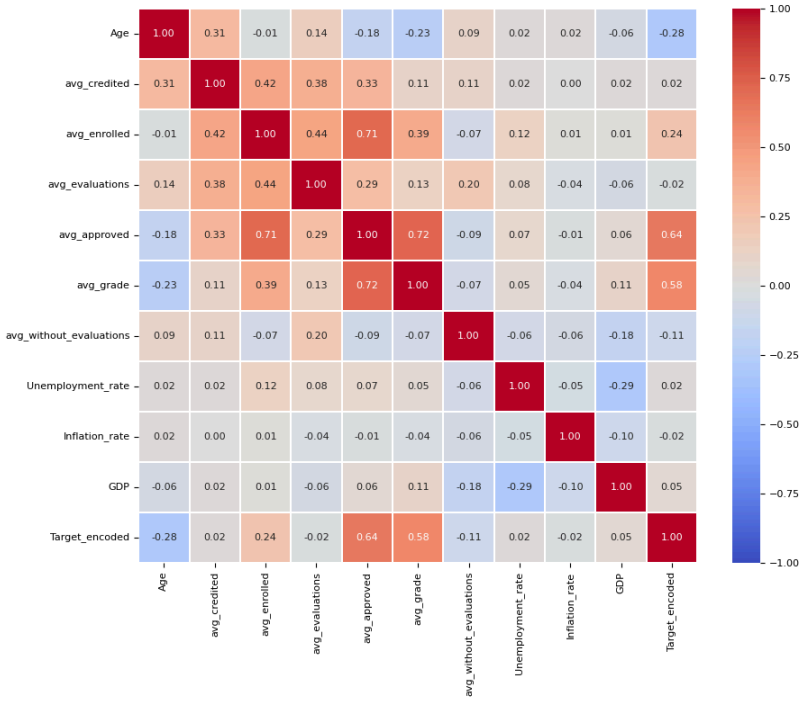
these have relatively high and positive correlations with the label, while some have very low correlations(e.g., 'Unemployment_rate', 'Inflation_rate')

The heat map also reveals multicollinearity among the features related to curricular units. These features represent students' academic performance at the end of the first and second semesters.

```
# Averaging academic performance data across two semesters
stud_selected['avg_credited'] = stud_selected[['Curricular_units_1st_sem_credited', 'Curricular_units_2nd_sem_credited']].mean(axis = 1)
stud_selected['avg_enrolled'] = stud_selected[['Curricular_units_1st_sem_enrolled', 'Curricular_units_2nd_sem_enrolled']].mean(axis = 1)
stud_selected['avg_evaluations'] = stud_selected[['Curricular_units_1st_sem_evaluations','Curricular_units_2nd_sem_evaluations']].mean(ax
stud_selected['avg_approved'] = stud_selected[['Curricular_units_1st_sem_approved','Curricular_units_2nd_sem_approved']].mean(axis = 1)
stud_selected['avg_grade'] = stud_selected[['Curricular_units_1st_sem_grade','Curricular_units_2nd_sem_grade']].mean(axis = 1)
stud_selected['avg_without_evaluations'] = stud_selected[['Curricular_units_1st_sem_without_evaluations', 'Curricular_units_2nd_sem_witho
```

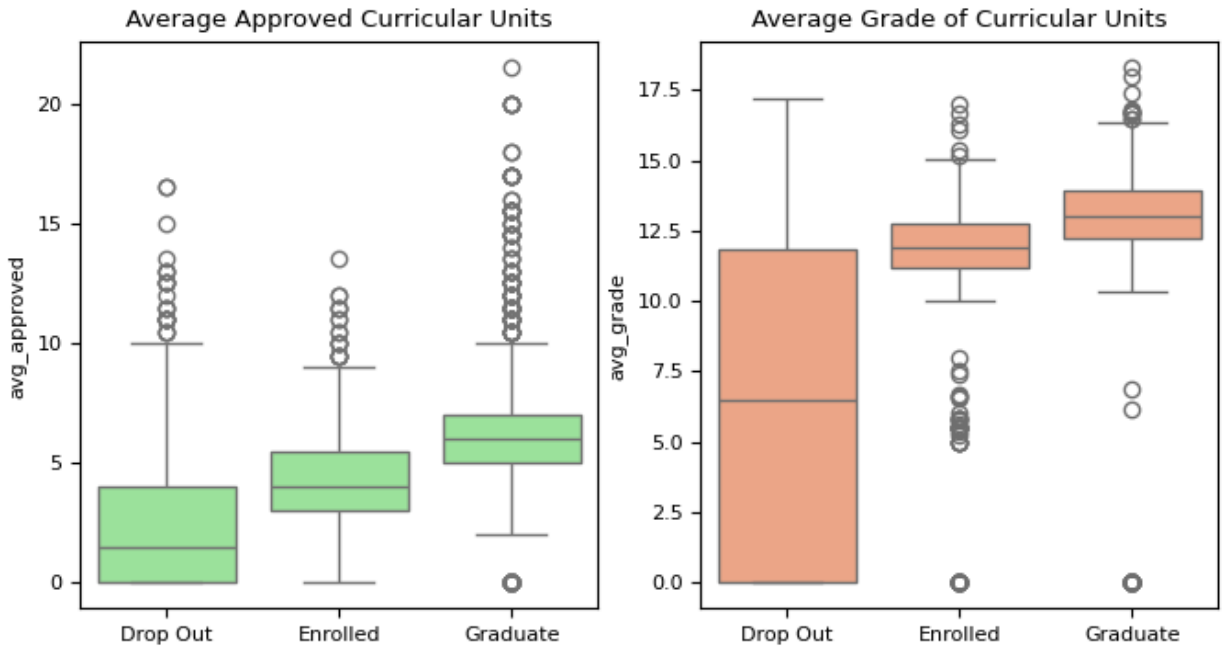
Python

```
# plot the heat map of correlation matrix again
num_features = stud_selected[['Age', 'avg_credited', 'avg_enrolled',
                              'avg_evaluations', 'avg_approved',
                              'avg_grade', 'avg_without_evaluations',
                              'Unemployment_rate', 'Inflation_rate',
                              'GDP', 'Target_encoded']]

plt.figure(figsize = (10, 8))
plt.rcParams.update({'font.size': 8})
sns.heatmap(num_features.corr(method = 'spearman'), vmin = -1, vmax = 1, center = 0,
            cmap = 'coolwarm', fmt = '.2f', linewidths = .2, annot = True);
```



The new correlation matrix above shows that 'curri_avg_approved' and 'curri_avg_grade' still have a relatively high correlation with the labels ('Target_encoded'), while 'curri_avg_credited' and 'curri_avg_evaluations', along with 'the macroeconomic data ('Unemployment_rate', 'Inflation_rate'), have very low correlations, all between -0.02 and 0.02.

```
# Plot 'avg_approved' and 'avg_grade' vs. Target
fig, (ax1, ax2) = plt.subplots(nrows = 1, ncols = 2, figsize = (8, 4))
sns.boxplot(data = stud_selected, x = 'Target_encoded', y = 'avg_approved',
            color = 'lightgreen', ax = ax1)
ax1.set_title('Average Approved Curricular Units')
ax1.set_xlabel("")
ax1.set_xticks([0, 1, 2])
ax1.set_xticklabels(['Drop Out', 'Enrolled', 'Graduate']);

sns.boxplot(data = stud_selected, x = 'Target_encoded', y = 'avg_grade',
            color = 'lightsalmon', ax = ax2)
ax2.set_title('Average Grade of Curricular Units')
ax2.set_xlabel("")
ax2.set_xticks([0, 1, 2])
ax2.set_xticklabels(['Drop Out', 'Enrolled', 'Graduate']);
```



V. Model and Techniques

A. Random Forests Model

Random forests were introduced in 2001 by Breiman which have emerged as a powerful ensemble learning material. Their accuracy and robustness makes them valuable in data

analysis. Despite the success in classification and regression tasks, the theoretical foundation regarding this remains vague. This paper discusses the fundamental mathematical properties and offers a theoretical analysis of random forest. The author shows that the random forest model is consistent and able to adapt to sparsity, in the sense that the rate of convergence depends only on the number of strong features rather than the full ambient dimension.

B. XGBoost Model

XGBoost is an open-source machine learning library that utilizes a gradient boosting framework to create a predicting model. The leveraging techniques and cache optimization that it possesses allows it to handle large datasets with efficiency. It also has a built-in regularization which helps prevent the model from overfitting.

C. Support Vector Machine Model

A support vector machine is supervised machine learning that is used for classifying data points into 2 distinct classes. The model identifies the maximum margin between the two classifications by placing a hyperplane that acts as a boundary to distinguish the classes. They can also be applied for non-linear data by applying kernel functions like polynomials and RBF. It is rarely affected by outliers as they focus on finding the furthest distance to divide the data points into 2 classifications.

D. Decision Tree

Decision tree is technique used for classification and regression tasks, it is characterized by their tree-like structure, Each tree consists of a node which represents features of a dataset, branches that represent outcomes and leaf nodes that represent final class labels. The tree is constructed by recursively partitioning the data based on feature values that minimize impurity making them versatile.

E. K-Nearest Neighbor

K-Nearest neighbor (KNN) is a straightforward machine learning algorithm used for classification and regression tasks. It works on the principle of finding the nearest ‘k’

data point in the feature space to an input, which makes predictions based on the neighbours(k). Distance between these points are usually measured using metrics such as Euclidean distance. KNN is non-parametric, which means it doesn't assume any underlying data distribution and it naturally adapts to the local structure of the data.

VI. Evaluation Techniques

A. Balanced Accuracy

Balanced accuracy is useful for imbalanced datasets, where one class may significantly outnumber the other. It calculates the average of the recall obtained on each class. It accounts for both true positive and true negative rates and provides a more representative measure of performance across all classes.

B. F1 Score

F1 Score is a machine learning metric that calculates precision and recall in order to evaluate a model's accuracy. This metric is commonly used in classification tasks. The formula of F1 Score is:

Precision = TP / (TP+FP)

C. AUC

Area Under the Curve (AUC) Score is a metric that measures how well a model can distinguish between positive and negative classes. It represents the probability that a model will rank a randomly chosen positive example higher than a randomly chosen negative example.

The AUC score ranges from 0 to 1, where:

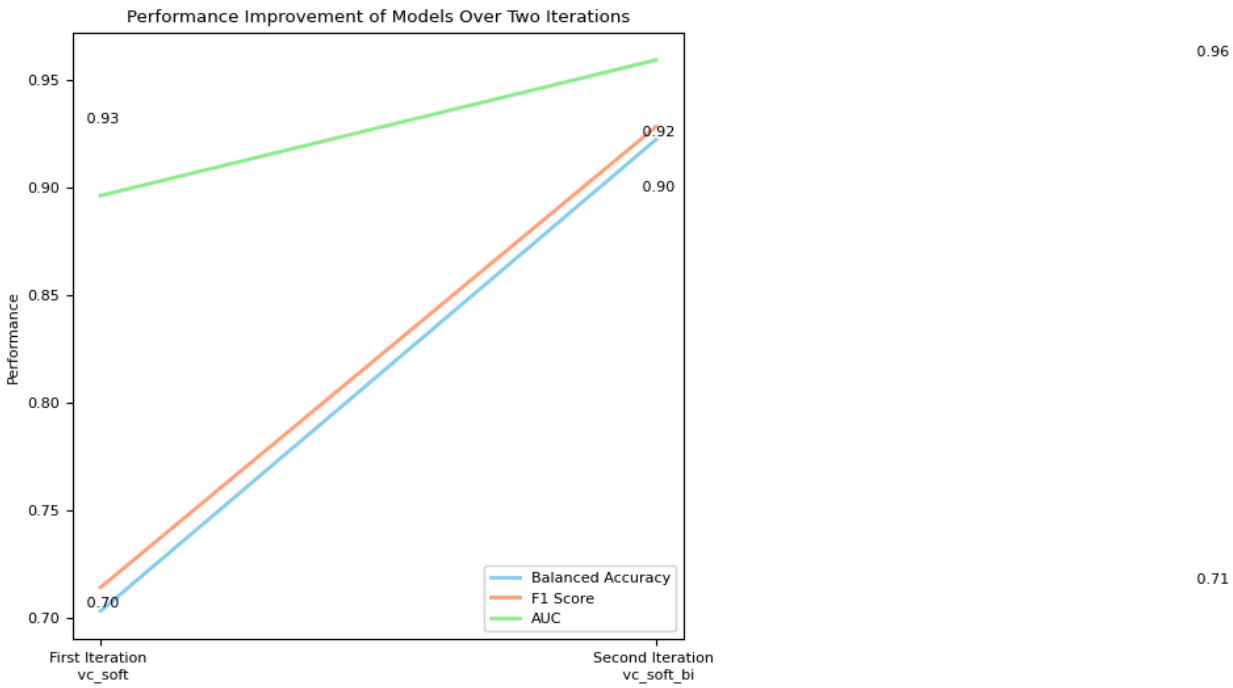
AUC = 1: The model can distinguish between positive and negative classes.

AUC < 0.5 : The model incorrectly ranks the negative classes higher than positive classes.

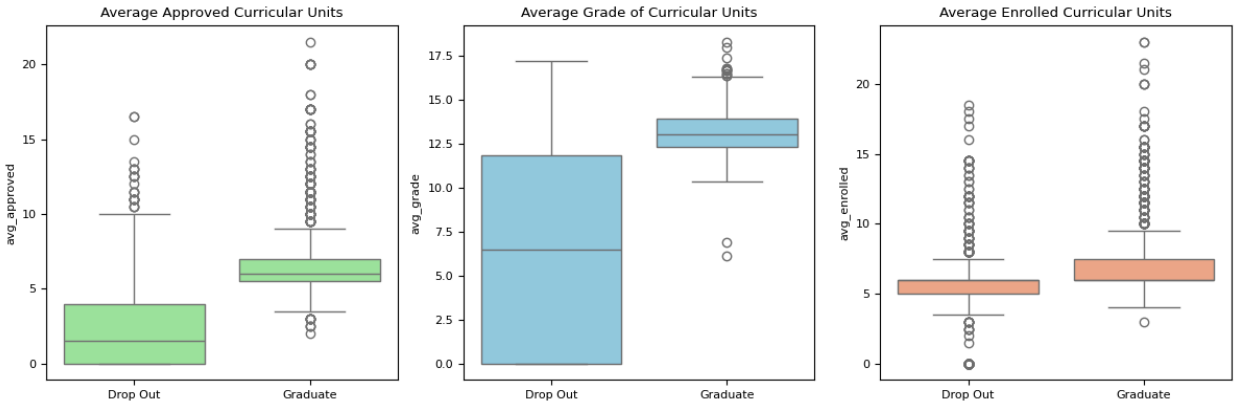
VII. Results and Discussions

A. Exploratory Data Analysis

This section delves into the performance of each model used within this project, along with understanding the correlation between the student attributes with dropout rates.



Between the first and second iteration of model testing, we removed rows containing the value “enrolled” in the Target attribute, thus only resulting in the values “Graduate” and “Dropout”. This method helps increase accuracy in the model by providing lesser outcomes for the prediction. As seen in the graph, there seems to be a major spike in accuracy between the first and second integration.



The graph gives us some insight regarding the trend of academic engagement within the university, an attribute that is important in understanding the patterns that spur students into dropping out or graduating. The attributes involved are:

1. Avg_approved: An attribute that represents the average number of circular units a student has completed during their enrollment.
2. Avg_grade: An attribute that represents the average grade obtained by the students across their courses.
3. Avg_enrolled: An attribute that represents the average number of circular units that a student is enrolled in during their academic year.

Box Plot no. 1 Analysis

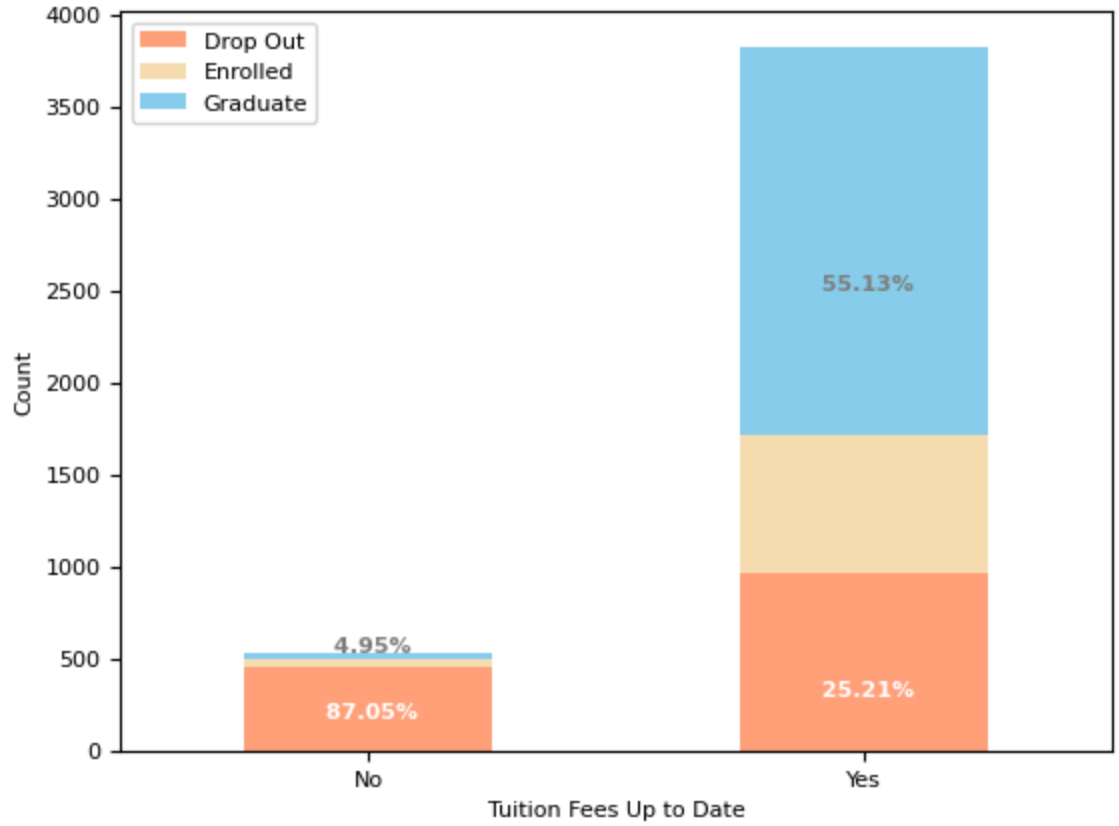
The graph shows that students who successfully complete more courses are more likely to graduate within their university life. Whereas students with fewer completed courses have higher tendencies of dropping out. Thus, showing that progression in academic accomplishments is an important factor of graduating.

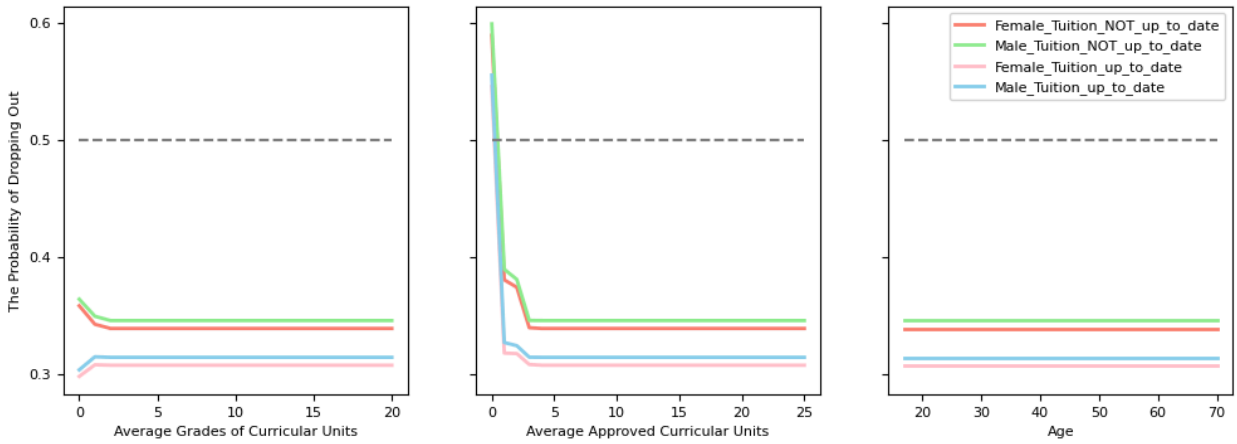
Box Plot no. 2 Analysis

Typically, grades are a common predictor on whether or not a student will graduate in most academic institutions. From the dataset, it shows that students with lower grades may not meet the academic requirements within the university, thus resulting in them to fail and drop out.

Box Plot no. 3 Analysis

Some majors may contain different numbers of courses. Meaning that the number of curricular units taken may not directly correlate to a student's academic engagement. This understanding provides another hypothesis that states that higher academic engagement may not conclude the choice of dropping out or graduating.



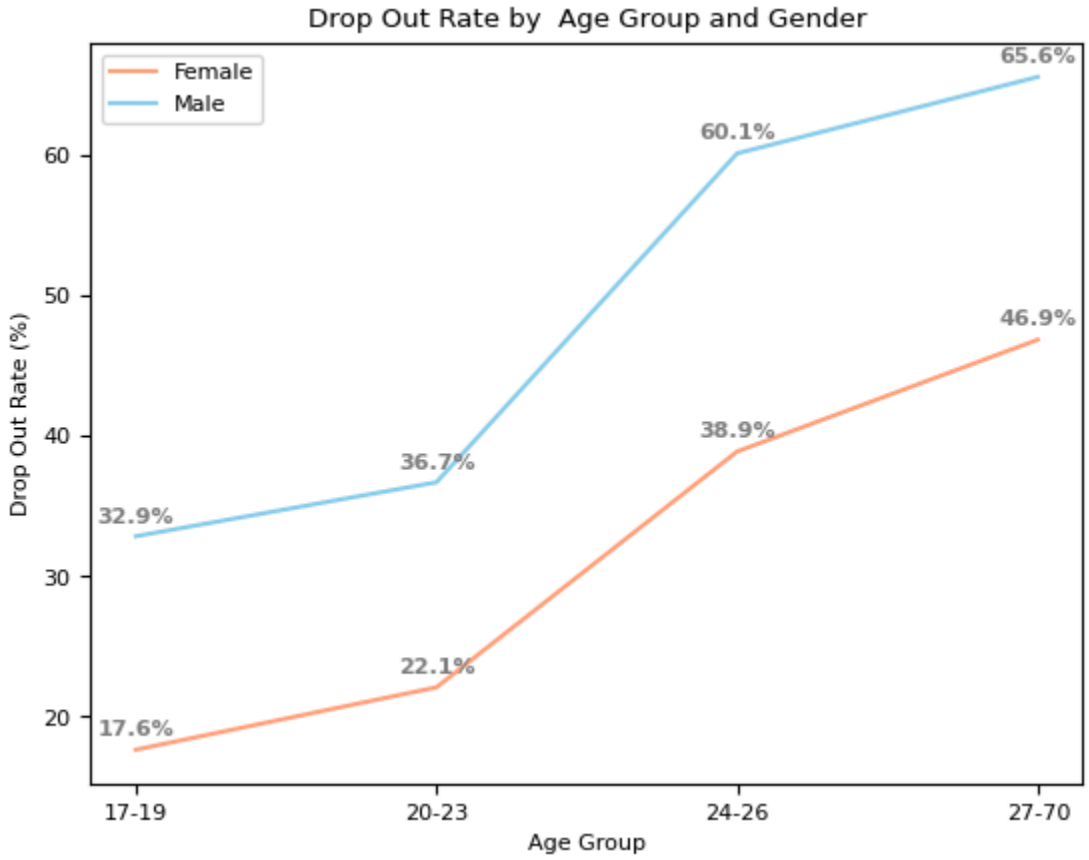


Academic accomplishments may not always be the main factor in university accomplishments. In some cases, it may resort to external factors like finance in addition to other attributes. The analysis for the graph displaying the contribution of 2 attributes are as follows:

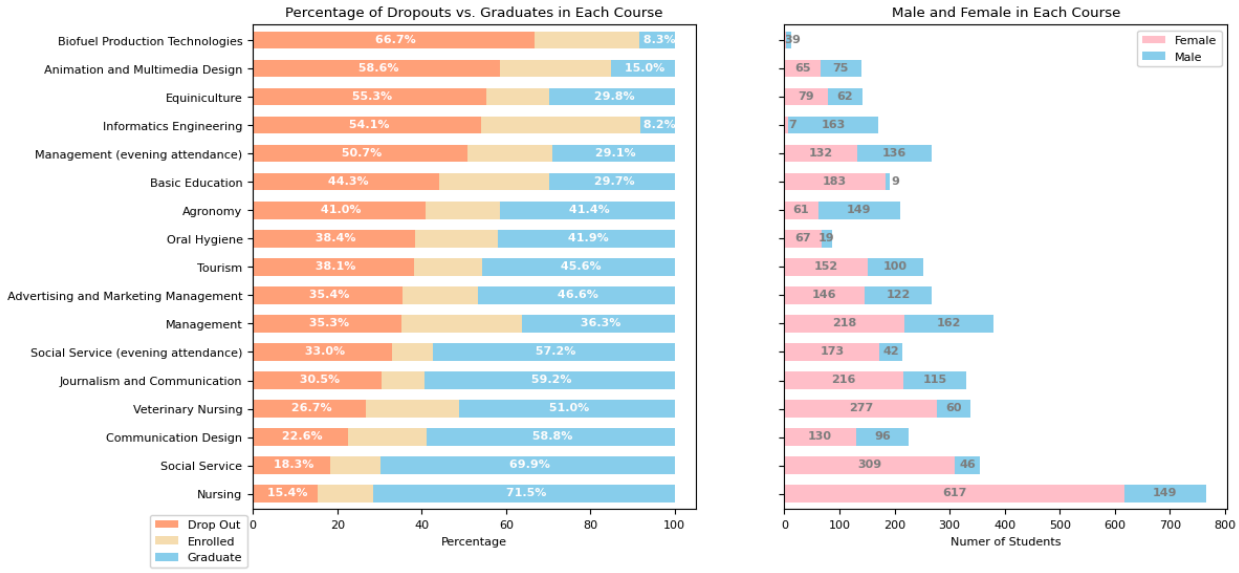
1. Academic performance is already a strong factor that leads to drop out rates, and financial instability simply amplifies the process of that. Perhaps universities may

offer financial support to those who are struggling academically due to financial issues.

2. By identifying the rates of accomplishment for each course, institutions can identify where they should provide educational guidance through the academic advisors and teacher associates.
3. Different age groups consist of different personality traits, younger students may drop out due to a lack of maturity whereas older students have larger responsibilities like supporting their family which may increase the rates of dropping out.



The age group seems to be directly proportional to the dropout rates of a student, adding another attribute into the conclusion. There seems to be a major spike between the second and third age group, displaying a drastic increase of 23.4% for male students and 16.8% for the female students.



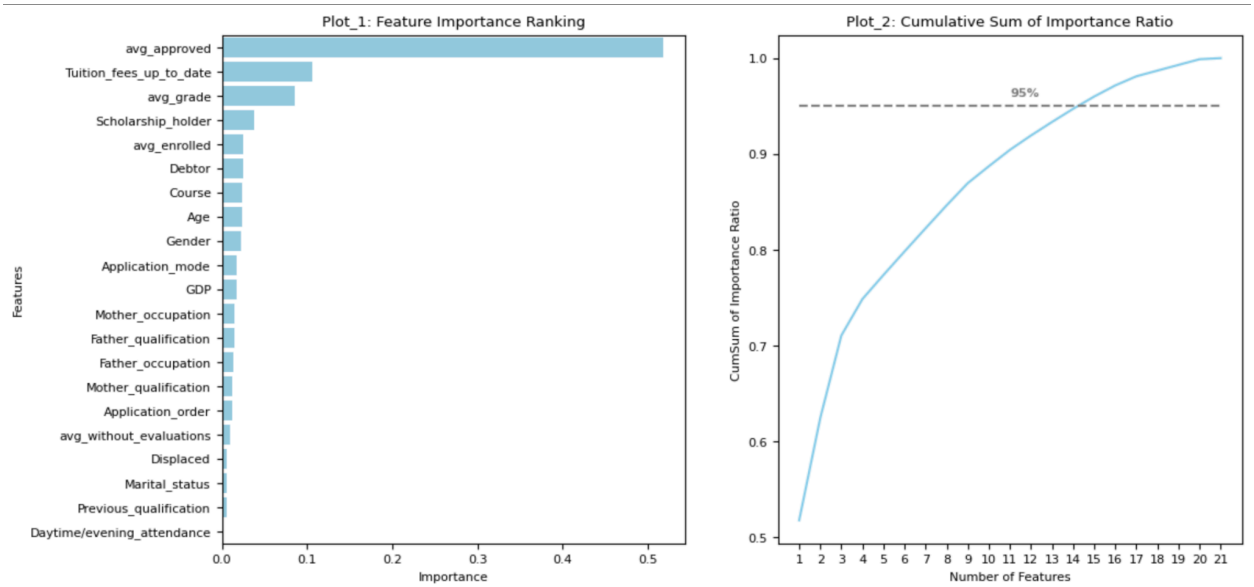
It is commonly known that different courses experience different levels of stress and difficulty. From the graph, we can identify that Biofuel Production Technologies and Animation are courses that lead to more drop out rates. Universities may use this as a reference to perhaps review the difficulty of material within the course.

VIII. Predictive Modelling

no.	Model	Balanced Accuracy	F1 Score	AUC
0	tuned_rf_bi	0.917	0.906	0.956
1	tuned_xgb_bi	0.921	0.908	0.962
2	svm_model_bi	0.891	0.899	0.940
3	tuned_dt_bi	0.894	0.894	0.931
4	best_knn_bi	0.845	0.854	0.907
5	vc_soft_bi	0.922	0.928	0.959

The results show that tuned XGB produced a high balanced accuracy with high F1 Score and the highest AUC score.

IX. Feature Importance



X. Conclusion and Future Works

This project aims to predict student dropout and graduate rates in order to identify the main contributing factors of why students finish or drop out of university. There are several attributes that we use as the factors. We utilized 4 machine learning models for the prediction, in the end, the XGBoost model resulted in the most accurate prediction.

We can conclude that the factors that mainly affect the dropout rates are course difficulty, finance, and gender. As for the graduation rate, it was mainly affected by the courses completion, grades, and courses number.

In the future, we aim to expand our research by utilizing a larger dataset with broader attributes and produce better accuracy by incorporating better machine learning models. We hoped that we can better understand the pattern of university student graduates and dropouts and help to increase student graduates and reduce student dropouts.

XI. Supplementary Codes

The dataset and the code for the models will be displayed in the following repository: https://github.com/Edvade/Data-Science-Final-Project.git

Dataset Source

https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

XII. References

[1] Lansford, J. E., Dodge, K. A., Pettit, G. S., & Bates, J. E. (2016). A Public Health Perspective on school dropout and adult Outcomes: A Prospective study of risk and protective factors from age 5 to 27 years. Journal of Adolescent Health, 58(6), 652–658. https://doi.org/10.1016/j.jadohealth.2016.01.014

[2]Ressa, T., & Andrews, A. (2022). High school dropout dilemma in America and the importance of reformation ofeducation systems to empower all students. International Journal of Modern Education Studies, 6(2), 423-447. https://doi.org/10.51383/ijonmes.2022.234

[3] Banaag, R. A., Sumodevilla, J. L., & Potane, J. D. (2024). Factors affecting student drop out behavior: A systematic review. *International Journal of Educational Management and Innovation*, *5*(1), 53-70. https://doi.org/10.12928/ijemi.v5i1.9396

[4] Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early detection of students at risk – Predicting student dropouts using administrative student data and machine learning methods. *SSRN Electronic Journal*, *11*(3). https://doi.org/10.2139/ssrn.3275433

[5]Amare, M. Y., & Simonova, S. (2021). Global challenges of students dropout: A prediction model development using machine learning algorithms on higher education datasets. *SHS Web of Conferences*, *129*, 09001. https://doi.org/10.1051/shsconf/202112909001

[6] Casanova, J., Cervero, A., Carlos Núñez, J., Almeida, L., & Bernardo, A. (2018). Factors that determine the persistence and dropout of university students. Psicothema, 30(4), 408–414. https://doi.org/10.7334/psicothema2018.155

[7] Francis Thaise Cimene, A Cimene, Albino, A.-A. C., & Villaflor. (2023, November 11). Understanding the Complex Factors behind Students Dropping Out of School. ResearchGate; unknown. https://www.researchgate.net/publication/375556929_Understanding_the_Complex_Factors_behind_Students_Dropping_Out_of_School