Decision Tree Analysis: Student's Dropout and Academic Success

Ryleigh Harvey and Natalie Dume

Dataset Description:

The dataset represents a compilation of data from higher education institutions, offering a rich source for synthesizing information, generating knowledge, and monitoring student progress.Instances of dropout and academic underachievement in higher education pose significant barriers to economic growth, employment, competitiveness, and productivity. These challenges directly impact students, their families, educational institutions, and society at large. The dataset combines information from various sources, encompassing demographic, socioeconomic, macroeconomic, and academic indicators of enrollment and performance across the first and second semesters. It serves as a foundational resource for constructing machine learning models to forecast academic outcomes and predict dropout rates.This dataset is invaluable for researchers seeking to conduct comparative analyses of student performance and for those engaged in machine learning training initiatives.

The values in the dataset represent:

Marital status: 1— Single 2— Married 3— Widower 4— Divorced 5— Facto union 6— Legally separated

Nationality: 1—Portuguese 2—German 3—Spanish 4—Italian 5—Dutch 6—English 7—Lithuanian 8—Angolan 9—Cape Verdean 10—Guinean 11—Mozambican 12—Santomean 13—Turkish 14—Brazilian 15—Romanian 16—Moldova (Republic of) 17—Mexican 18—Ukrainian 19—Russian 20—Cuban 21—Colombian

Application mode: 1—1st phase—general contingent 2—Ordinance No. 612/93 3—1st phase—special contingent (Azores Island) 4—Holders of other higher courses 5—Ordinance No. 854-B/99 6—International student (bachelor) 7—1st phase—special contingent (Madeira Island) 8—2nd phase—general contingent 9—3rd phase—general contingent 10—Ordinance No. 533-A/99, item b2) (Different Plan) 11—Ordinance No. 533-A/99, item b3 (Other Institution) 12—Over 23 years old 13—Transfer 14—Change in course 15—Technological specialization diploma holders 16—Change in institution/course 17—Short cycle diploma holders 18—Change in institution/course (International)

Course values: 1—Biofuel Production Technologies 2—Animation and Multimedia Design 3—Social Service (evening attendance) 4—Agronomy 5—Communication Design 6—Veterinary Nursing 7—Informatics Engineering 8—Equiniculture 9—Management 10—Social Service 11—Tourism 12—Nursing 13—Oral Hygiene 14—Advertising and Marketing Management 15—Journalism and Communication 16—Basic Education 17—Management (evening attendance)

Previous qualification: 1—Secondary education 2—Higher education—bachelor's degree 3—Higher education—degree 4—Higher education—master's degree 5—Higher education—doctorate 6—Frequency of higher education 7—12th year of schooling—not completed 8—11th year of schooling—not completed 9—Other—11th year of schooling 10—10th year of schooling 11—10th year of schooling—not completed 12—Basic education 3rd cycle (9th/10th/11th year) or equivalent 13—Basic education 2nd cycle (6th/7th/8th year) or equivalent 14—Technological specialization course 15—Higher education—degree (1st cycle) 16—Professional higher technical course 17—Higher education—master's degree (2nd cycle)

Mother's and Father's qualification: 1—Secondary Education—12th Year of Schooling or Equivalent 2—Higher Education—bachelor’s degree 3—Higher Education—degree 4—Higher Education—master’s degree 5—Higher Education—doctorate 6—Frequency of Higher Education 7—12th Year of Schooling—not completed 8—11th Year of Schooling—not completed 9—7th Year (Old) 10—Other—11th Year of Schooling 11—2nd year complementary high school course 12—10th Year of Schooling 13—General commerce course 14—Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent 15—Complementary High School Course 16—Technical-professional course 17—Complementary High School Course—not concluded 18—7th year of schooling 19—2nd cycle of the general high school course 20—9th Year of Schooling—not completed 21—8th year of schooling 22—General Course of Administration and Commerce 23—Supplementary Accounting and Administration 24—Unknown 25—Cannot read or write 26—Can read without having a 4th year of schooling 27—Basic education 1st cycle (4th/5th year) or equivalent 28—Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent 29—Technological specialization course 30—Higher education—degree (1st cycle) 31—Specialized higher studies course 32—Professional higher technical course 33—Higher Education—master’s degree (2nd cycle) 34—Higher Education—doctorate (3rd cycle)

Mother's and Father's occupation: 1—Student 2—Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 3—Specialists in Intellectual and Scientific Activities 4—Intermediate Level Technicians and Professions 5—Administrative staff 6—Personal Services, Security and Safety Workers, and Sellers 7—Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry 8—Skilled Workers in Industry, Construction, and Craftsmen 9—Installation and Machine Operators and Assembly Workers 10—Unskilled Workers 11—Armed Forces Professions 12—Other Situation; 13—(blank) 14—Armed Forces Officers 15—Armed Forces Sergeants 16—Other Armed Forces personnel 17—Directors of administrative and commercial services 18—Hotel, catering, trade, and other services directors 19—Specialists in the physical sciences, mathematics, engineering, and related techniques 20—Health professionals 21—Teachers 22—Specialists in finance, accounting, administrative organization, and public and commercial relations 23—Intermediate level science and engineering technicians and professions 24—Technicians and professionals of intermediate level of health 25—Intermediate level technicians from legal, social, sports, cultural, and similar services 26—Information and communication technology technicians 27—Office workers, secretaries in general, and data processing operators 28—Data, accounting, statistical, financial services, and registry-related operators 29—Other administrative support staff 30—Personal service workers 31—Sellers 32—Personal care workers and the like 33—Protection and security services personnel 34—Market-oriented farmers and skilled agricultural and animal production workers 35—Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence 36—Skilled construction workers and the like, except electricians 37—Skilled workers in metallurgy, metalworking, and similar 38—Skilled workers in electricity and electronics 39—Workers in food processing, woodworking, and clothing and other industries and crafts 40—Fixed plant and machine operators 41—Assembly workers 42—Vehicle drivers and mobile equipment operators 43—Unskilled workers in agriculture, animal production, and fisheries and forestry 44—Unskilled workers in extractive industry, construction, manufacturing, and transport 45—Meal preparation assistants 46—Street vendors (except food) and street service providers

Gender: 1—male 0—female

Daytime/evening attendance: 1—daytime 0—evening

Displaced, Educational special needs, Debtor, Tuition fees up to date, Scholarship holder, International:

1—yes 0—no

In [1]:
```
#import necessary libraries
import numpy as np
import pandas as pd
import pydotplus
from sklearn.tree import DecisionTreeClassifier
from sklearn import datasets
from IPython.display import Image
from sklearn import tree
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectFromModel
import matplotlib.pyplot as plt
```

In [2]:
```
df = pd.read_csv("data_sets/data.csv", delimiter=';') # loading dataset
```

In [3]:
```
df.head() #display first few rows of dataset
```

Out[3]:

	Marital status	Application mode	Application order	Course	Daytime/evening attendance\t	Previous qualification	Previous qualification (grade)	Na
0	1	17	5	171	1	1	122.0	
1	1	15	1	9254	1	1	160.0	
2	1	1	5	9070	1	1	122.0	
3	1	17	2	9773	1	1	122.0	
4	2	39	1	8014	0	1	100.0	

5 rows × 37 columns

In [4]:
```
#Define features and target variables
features = ['Marital status', 'Application mode', 'Application order', 'Cour
       'Daytime/evening attendance\t', 'Previous qualification',
       'Previous qualification (grade)', 'Nacionality',
       "Mother's qualification", "Father's qualification",
       "Mother's occupation", "Father's occupation", 'Admission grade',
       'Displaced', 'Educational special needs', 'Debtor',
       'Tuition fees up to date', 'Gender', 'Scholarship holder',
       'Age at enrollment', 'International',
       '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']

x = df[features] #features
y = df.Target  #target variable
class_names = list(y.unique())
```

In [5]:
```
#create decision tree
decisiontree = DecisionTreeClassifier(random_state=0)
```

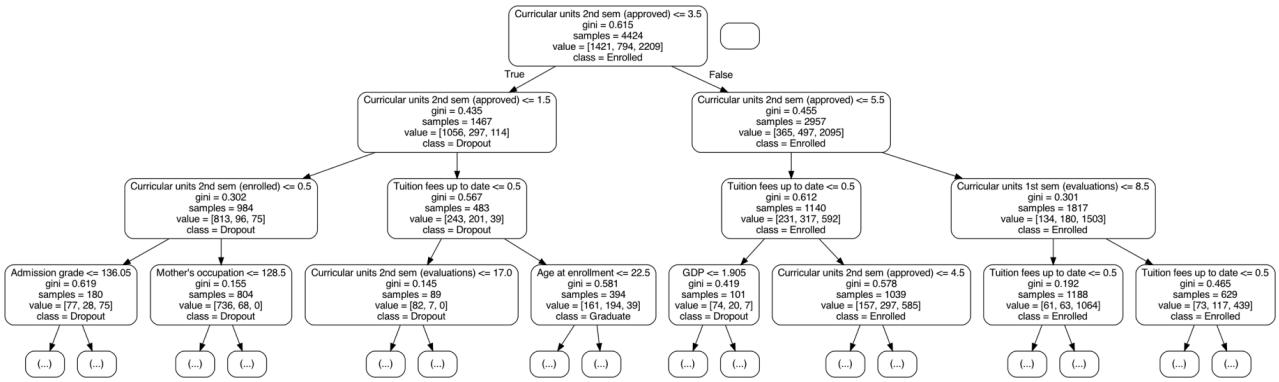
In [6]:
```
#fit the model
model = decisiontree.fit(x, y)
```

In [7]:
```
#export visualization
dot_data = tree.export_graphviz(model,
                                out_file=None,
                                 filled=False,
                                 rounded=True,
                                 max_depth=3,
                                 special_characters=False,
                                feature_names=features,class_names= class_nam
```

In [8]:
```
#create graph
graph = pydotplus.graph_from_dot_data(dot_data)
```

In [9]:
```
#display tree
Image(graph.create_png())
```

Out[9]:



This decision tree model was constructed using the entire dataset and subsequently examined various categories within the data to analyze their effects on the target variable. These categories are seperated into Demographic, Socioeconomic, and Macroeconomic.

demographic data

The decision tree splits based on demographic features such as marital status, nationality, displacement status, gender, age at enrollment, and international status. This suggests that these demographic factors play a significant role in predicting the target variable.

In [10]:
```
#features for demographic tree
features = ['Marital status','Nacionality','Displaced','Gender','Age at enrc
x = df[features]
y = df.Target
class_names = list(y.unique())
```

In [11]:
```
decisiontree = DecisionTreeClassifier(random_state=0)
```

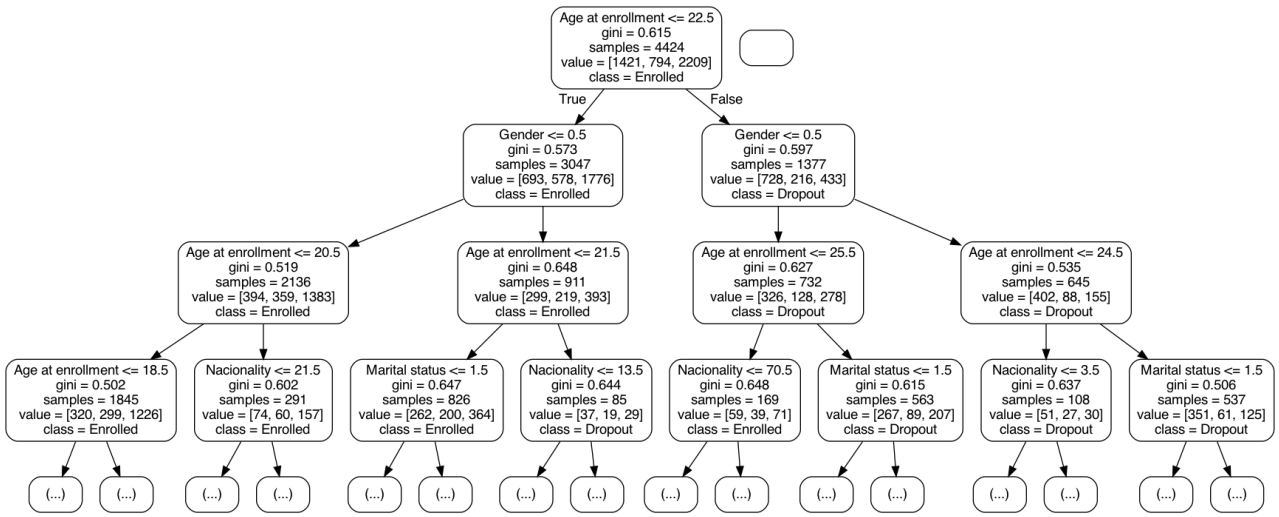
In [12]:
```
model = decisiontree.fit(x, y)
```

In [13]:
```
dot_data = tree.export_graphviz(model,
                                out_file=None,
                                 filled=False,
                                 rounded=True,
                                 max_depth =3,
                                 special_characters=False,
                                feature_names=features,class_names= class_nam
```

In [14]:
```
graph1 = pydotplus.graph_from_dot_data(dot_data)
```

In [15]:
```
Image(graph1.create_png())
```

Out[15]:



socioeconomics

The decision tree further splits based on socioeconomic factors such as parents' qualifications, parents' occupation, educational special needs, debtor status, tuition fee payment status, and scholarship status. This indicates that socioeconomic factors also contribute to predicting the target variable.

In [16]:
```
#features for socioeconomic tree
features = ["Father's qualification","Mother's qualification","Father's occu
x = df[features]
y = df.Target
class_names = list(y.unique())
```

In [17]:
```
decisiontree = DecisionTreeClassifier(random_state=0)
```

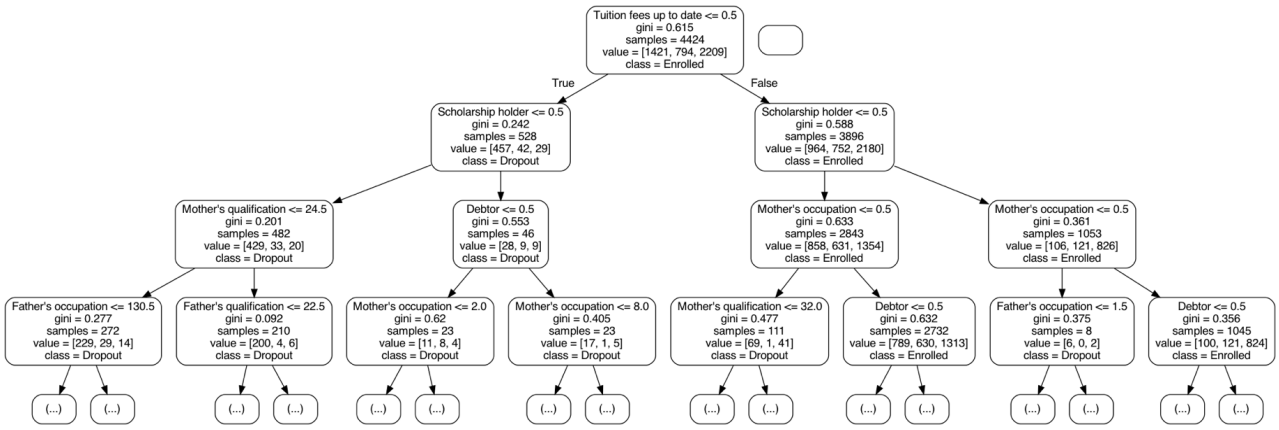
In [18]:
```
model = decisiontree.fit(x, y)
```

In [19]:
```
dot_data = tree.export_graphviz(model,
                                out_file=None,
                                 filled=False,
                                 rounded=True,
                                 max_depth =3,
                                 special_characters=False,
                                feature_names=features,class_names= class_nam
```

In [20]:
```
graph2 = pydotplus.graph_from_dot_data(dot_data)
```

In [21]:
```
Image(graph2.create_png())
```

Out[21]:



macroeconomic

The decision tree branches based on macroeconomic indicators such as unemployment rate, inflation rate, and GDP.These variables appear to have less influence compared to demographic and socioeconomic factors, as the tree depth is deeper for these features.

In [22]:
```
#features for macroeconomic tree
features = ["Unemployment rate","Inflation rate","GDP"]
x = df[features]
y = df.Target
class_names = list(y.unique())
```

In [23]:
```
decisiontree = DecisionTreeClassifier(random_state=0)
```

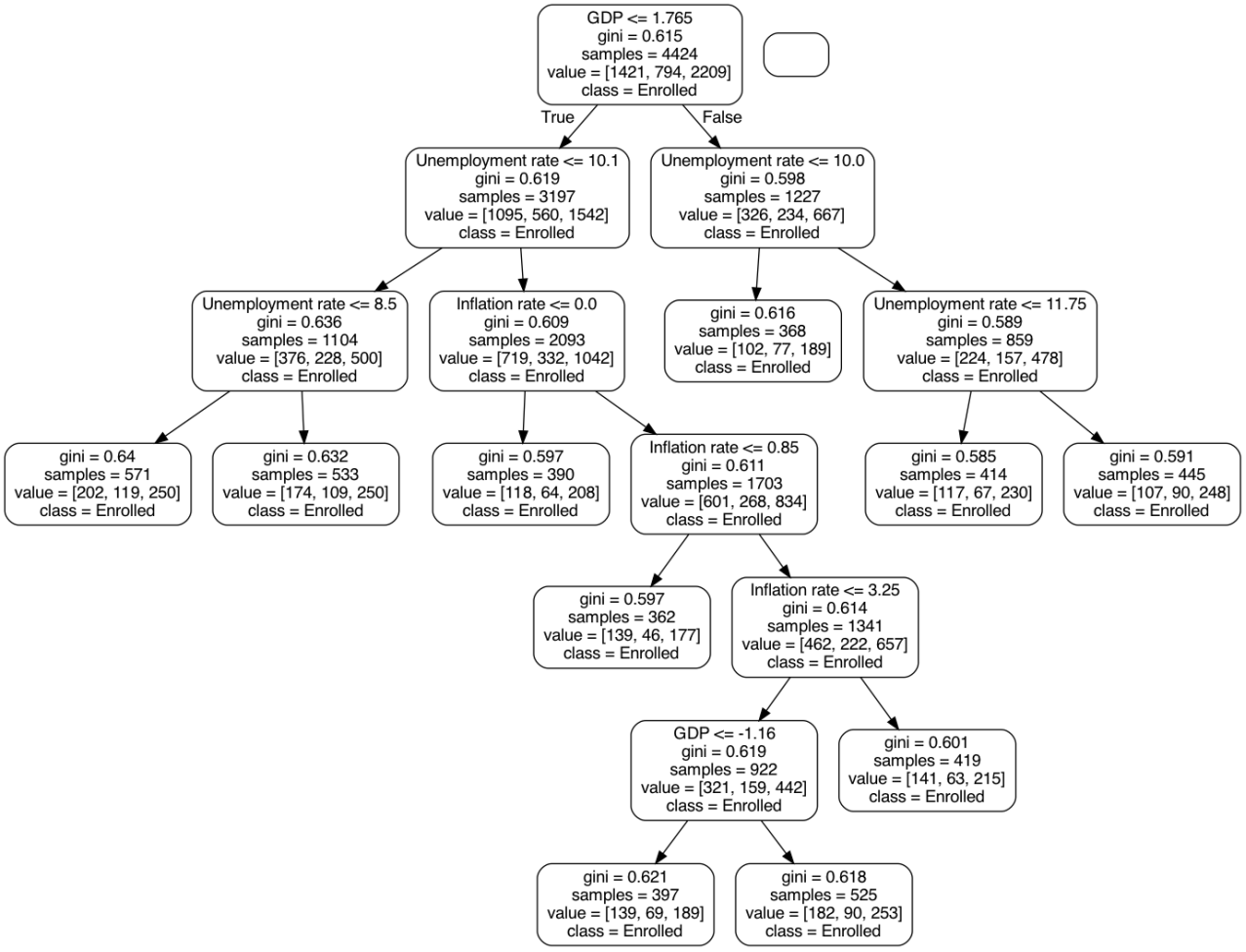
In [24]:
```
model = decisiontree.fit(x, y)
```

In [25]:
```
dot_data = tree.export_graphviz(model,
                                out_file=None,
                                 filled=False,
                                 rounded=True,
                                 max_depth =10,
                                 special_characters=False,
                                feature_names=features,class_names= class_nam
```

In [26]:
```
graph3 = pydotplus.graph_from_dot_data(dot_data)
```

In [27]:
```
Image(graph3.create_png())
```

Out[27]:



Conclusion

The decision tree models demonstrates the importances of various factors in predicting the target variable.The decision tree analysis provides valuable insights into the factors influencing student academic performance or dropout rates in higher education. By examining demographic, socioeconomic, and macroeconomic variables, we gain a multifaceted understanding of the complex dynamics at play. Demographic factors such as martial status, nationality, and age are significant precidictors, suggesting that individual caharcteristics play a crucialmrole in shaping studnet outcomes. Socioeconomic indicators like parents' qualifications, occupation, and financial status also exert considerable influence, highlighting the importance of socioeconomic background in educational attainment.Demographic and socioeconomic factors seem to have a more significant impact on the target variable compared to macroeconomic indicators. This underscores the intricate interplay between personal circumstances and broader economic conditions in shaping educational trajectories.Overall, the decision tree visualizations provide a nuanced understanding of the predictors of academic success and dropout in higher education. These insights can inform targeted interventions and policies aimed at improving outcomes for students from diverse backgrounds, ultimately fostering greater equity and inclusivity within higher education systems.

Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.

In [ ]: