**Investigating the Influence of Social and Economic Factors on Student Dropout and Academic Achievement**

**Project report**

**Group 10**

**Rana Hamdan**

**Yashashri Kadam**

**Sree Sai Chitra Duvvapu**

**Introduction**

Higher education institutions hold extensive student data, providing ample opportunities for generating insights, enriching knowledge, and facilitating efficient monitoring. Academic achievement in higher education plays a pivotal role in advancing society, creating employment opportunities, and fostering economic prosperity. Yet, dropout rates and educational setbacks pose significant challenges to these goals, exerting direct repercussions on students, families, institutions, and broader society. Effectively addressing these challenges requires a comprehensive examination of dropout trends and the various factors contributing to them. This entails an investigation into the impact of social and economic variables on student dropout.

**Data Collection**

The dataset we are using was downloaded from the Kaggle site available at:

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

The dataset provides a comprehensive view of students enrolled in various undergraduate degrees offered at a higher education institution. It includes demographic data, socioeconomic factors, and academic performance information that can be used to analyze the possible predictors of student dropout and academic success. This data can be used to estimate overall student performance at the end of each semester by assessing curricular units credited/enrolled/evaluated/approved as well as their respective grades. Finally, we have the unemployment rate, inflation rate, and GDP from the region which can help us further understand how economic factors play into student dropout rates or academic success outcomes. This powerful analysis tool will provide valuable insight into what motivates students to stay in school or abandon their studies for a wide range of disciplines such as agronomy, design, education nursing journalism management social service, or technologies.

The dataset contains student records from the academic years 2008/2009 to 2018/2019 across 17 undergraduate degrees in diverse fields such as agronomy, design, education, nursing, journalism, management, social service, and technologies. It comprises 4424 records with 35 attributes, in a CSV file. Notably, the dataset contains no missing values, ensuring robustness for analysis and modelling tasks.

| **Variable** | **Description** |
| --- | --- |
| Marital\_status | Single (1), Married (2) other (3) (categorical) |
| Application\_mode | Regular Application Phases(1),Special Cases and Exceptions(2) Specific Ordinances and Qualifications(3) / (categorical) |
| Application\_order | The order in which the student applied(Numerical) |
| Course | The course is taken by the student. Technology and Engineering(1),Design and Media(2),Healthcare and Nursing(3),Management and Social Services(4),Other(5)(Categorical) |
| Daytime\_evening\_attendance | Whether the student attends classes during the day or in the evening. (Categorical) daytime(1),  evening(0) |
| Previous\_qualification | The qualification obtained by the student before enrolling in higher education. (Categorical), Basic Education(1), Higher Education(2) |
| Nationality | The nationality of the student.Portuguese(1 ),Not Portuguese(0) |
| Mother\_qualification | The qualification of the student's mother.Basic Education(1), Higher Education(2),Others(3) |
| Father\_qualification | The qualification of the student's father.Basic Education(1), Higher Education(2),Others(3) |
| Mother\_occupation | The occupation of the student's mother. Professional and Managerial Occupations(1), Technical and Skilled Trades (2), Others(3) |
| Father\_occupation | The occupation of the student's father. Professional and Managerial Occupations(1), Technical and Skilled Trades (2), Others(3) |
| Displaced | Whether the student is a displaced person.yes(1), no (0) |
| Educational\_special\_needs | Whether the student has any special educational needs. Yes (1), No(0) |
| Debtor | Whether the student is a debtor. Yes (1), No (0) |
| Tuition\_fees\_are\_up\_to\_date | Whether the student's tuition fees are up to date. Yes(1), No(0) |
| Gender | The gender of the student. Male (1), female (0) |
| Scholarship\_holder | Whether the student is a scholarship holder, Yes(1), No(0) |
| Age\_at\_enrollment | The age of the student at the time of enrollment. |
| International | Whether the student is an international student. |
| Curri\_units\_1st\_sem\_credited | The number of curricular units credited by the student in the first semester. |
| Curri\_units\_1st\_sem\_enrolled | The number of curricular units enrolled by the student in the first semester. |
| Curri\_units\_1st\_sem\_evaluations | The number of curricular units evaluated by the student in the first semester. |
| Curri\_units\_1st\_sem\_approved | The number of curricular units approved by the student in the first semester. (Numerical) |
| Curri\_units\_1st\_sem\_grade | Curricular Units 1st semester grade |
| Curri\_units\_1st\_sem\_woeval | Curricular Units 1st semester (without Evaluations) |
| Curri\_units\_2st\_sem\_credited | The number of curricular units credited by the student in the second semester. |
| Curri\_units\_2st\_sem\_enrolled | The number of curricular units enrolled by the student in the second semester. |
| Curri\_units\_2st\_sem\_evaluations | The number of curricular units evaluated by the student in the second semester. |
| Curri\_units\_2st\_sem\_approved | The number of curricular units approved by the student in the second semester. (Numerical) |
| Curri\_units\_2st\_sem\_grade | Curricular Units second semester grade |
| Curri\_units\_2st\_sem\_woeval | Curricular Units second semester (without Evaluations) |
| Unemployment\_rate | Unemployment rate in the region |
| Inflation\_rate | Inflation rate from the region |
| GDP | Gross Domestic Product |
| Education\_Status | Dropout (1), non-dropout (0) |

**Data Cleaning / Dimension reduction**

Several variable names contained spaces, requiring modification to function properly in SAS. Additionally, certain categorical variables exhibited numerous levels, with some reaching up to 35 levels. After assessing the frequency of these levels, these adjustments were implemented:

Marital status was consolidated from 6 levels to 3, while Nationality was streamlined from 22 levels to 2. Application mode transformed 18 levels to 3, and the course variable was condensed from 17 levels to 5. Previous qualifications were simplified from 17 to 2 levels, and both mothers' and fathers' qualifications were streamlined from 29 and 35 levels to 3, respectively. Mother and father occupations were also streamlined from 46 and 32 levels to 3 each.

A screenshot of a data

Description automatically generated

A screenshot of a table

Description automatically generatedA screenshot of a graph

Description automatically generated

A table of numbers and a course

Description automatically generatedA screenshot of a graph

Description automatically generated

A screenshot of a table

Description automatically generatedA screenshot of a table

Description automatically generated

A screenshot of a table

Description automatically generated     A screenshot of a table

Description automatically generated

A table of numbers and numbers

Description automatically generated

**Data reduction:**

**A white rectangular box with black numbers

Description automatically generated A screenshot of a data

Description automatically generated**

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**A white text on a white background

Description automatically generatedA screenshot of a computer

Description automatically generated**

**A table of medical services

Description automatically generated with medium confidence A screenshot of a graph

Description automatically generated**

**A white paper with black text

Description automatically generated A screenshot of a computer

Description automatically generated**

**A white paper with black text

Description automatically generatedA screenshot of a computer

Description automatically generated**

**A white and black text on a white background

Description automatically generated A screenshot of a graph

Description automatically generated**

**Correlation matrix**

A screenshot of a computer

Description automatically generated

From the correlation matrix for identifying the correlation between the numerical variables which are 16 variables we could see there is a strong relationship between some of the variables :

* Curricular units 1st sem (credited) and Curricular units 2nd sem (credited) (0.9448).
* Curricular units 1st sem (enrolled) and Curricular units 2nd sem (enrolled) (0.9426).
* Curricular units 1st sem (approved) and Curricular units 2nd sem (approved) (0.9040).

**PCA**

A screenshot of a computer

Description automatically generated

A table of numbers with numbers on it

Description automatically generatedA graph of a plot

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

The minimum number of principal components needed to account for at least 80% of variability is 6. By observing the cumulative value in the below table, we can understand that the cumulative of 6 principle components is needed for at least 80% variability. Curri\_units\_1st\_sem\_enrolled (0.36)and Curri\_units\_1st\_sem\_approved (0.36) are the most important features in the first component. Curri\_units\_1st\_sem\_woeval (0.34) and Age\_at\_enrollment (0.32)are the most important features in the second component. Curri\_units\_2nd\_sem\_woeval (0.55) and Curri\_units\_1st\_sem\_woeval (0.54) are the most important features in the third component.

**Graphs and Summary Statistics (Exploratory Data Analysis)**

**A graph showing a dropout

Description automatically generatedA screenshot of a computer screen

Description automatically generated**

The outcome variable, Education\_Status, is a binary variable: 1 indicating dropout and 0 indicating non-dropout. The two classes show some imbalance. The non-dropout category, constituting 67.88% of the dataset, encompasses 3003 records out of 4424. Conversely, the dropout category represents 32% of the total records, with 1421 instances out of 4424.

A graph of age at enrollment

Description automatically generatedA screenshot of a computer screen

Description automatically generated

The mean age of students is 23, with ages ranging from 17 to 70, with a predominant concentration of values in the twenties.

**A graph of a graph showing the age and education status

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

The average age of dropout students (26.07) exceeds that of non-dropout students (21.07). Additionally, there is greater variability in age among dropout students compared to non-dropout students. Both distributions are positively skewed and exhibit numerous outliers. In the case of non-dropout students, ages above 30 are considered outliers, whereas for dropout students, the threshold is approximately 45.

The summary statistics of the macroeconomics data

**A screenshot of a graph

Description automatically generated**

The summary statistics of academic data at the end of the first semester:

**A screenshot of a computer

Description automatically generated**

The summary statistics of academic data at the end of the second semester:

**A screenshot of a computer

Description automatically generated**

**A graph showing a number of people

Description automatically generated with medium confidenceA screenshot of a table

Description automatically generated**

The variable “Gender” is a binary variable with female category 0 representing 64.83% while male 1 is 35.17 %.

**A graph of a graph showing a diagram

Description automatically generated with medium confidenceA graph of a diagram

Description automatically generated with medium confidence**

The grades for dropout students seem to have more variability than non-**drop**out students and as expected the grades for non-dropout have higher averages. There are a lot of outliers for non-dropout. drop out student's grades have clear negative skewness while non-dropout seems more symmetrical.

**Data Analysis**

The goal of our analysis is to determine whether a student will be classified as a dropout (1) or not dropout (0). To conduct this analysis, we have chosen three data mining techniques: Classification and Regression Tree (CART, neural network (NN), and logistic regression. Since our outcome is binary, these three techniques were chosen as they are used for classification. For each of these techniques, we built several models with different parameters to identify the most accurate model. To maintain consistency, each model split the dataset into 80% for the training set and 20% for the validation set and used the seed of 12345.

**Classification and Regression Tree (CART)**

| Model | Details |
| --- | --- |
| Model 1 | GINI CART model |
| Model 2 | GINI CART model with variable importance |

The first method we used for model building was CART where we built two different models, as summarized above. For one of the models, we used all the variables from our data set , using gini. The other model used only the important variables determined by CART, which were “Curri\_units\_2nd\_sem\_approved”, “tuition\_upToDate”, “Curri\_units\_1st\_sem\_enrolled”, “age\_at\_enrollment”, “curri\_units\_2nd\_sem\_grade”, “course”, “application order”, “Application\_order”, “Curri\_units\_2nd \_sem\_enrolled”, “Curri\_units\_2nd\_sem\_credited”, “Father\_occupation”, “Curri\_units\_1st\_sem\_enrolled”, “ Daytime\_evening\_attendance”, “Curri\_units\_2nd\_sem\_evaluations”, “inflation\_rate”, “Unemployment\_rate” and “Educational\_special\_needs”. To compare the models, we used the misclassification rate, sensitivity, specificity, and area under the curve (AUC).

|  | **Misclassification** | | **Sensitivity** | | **Specificity** | | **AUC** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Training** | **Validation** | **Training** | **Validation** | **Training** | **Validation** | **Training** | **Validation** |
| **Model 1** | 11.09 | 13.61 | 72.6 | 76.13 | 96.63 | 95.64 | 90.57 | 89.25 |
| **Model 2** | 11.43 | 12.93 | 73.04 | 71.33 | 97.89 | 94.63 | 90.36 | 89.25 |

As you can see from the results, Models 1 and 2 perform almost identically in both the training and validation sets with the AUC. Sensitivity higher on the validation set for model1 and higher on training set for model 2 and there is no indication of overfitting since there isn’t a huge gap between the training and validation sets in each model. The misclassification rate is similar on both the models with slight differences. In model2, the misclassification rate is less on the validation set than in the model1. Overall, we decided that Model 2 would be the best model since it has the lowest misclassification rate on validation set among both models.

**Neural Network (NN)**

| Model | Details |
| --- | --- |
| Model 1 | NN model |
| Model 2 | NN model with variable importance |

The second method we used for model building was Neural Network where we built two different models, as summarized above. For one of the models, we used all the variables from our data set. The other model used only the important variables determined by CART, which were “Curri\_units\_2nd\_sem\_approved”, “tuition\_upToDate”, “Curri\_units\_1st\_sem\_enrolled”, “age\_at\_enrollment”, “curri\_units\_2nd\_sem\_grade”, “course”, “application order”, “Application\_order”, “Curri\_units\_2nd \_sem\_enrolled”, “Curri\_units\_2nd\_sem\_credited”, “Father\_occupation”, “Curri\_units\_1st\_sem\_enrolled”, “ Daytime\_evening\_attendance”, “Curri\_units\_2nd\_sem\_evaluations”, “inflation\_rate”, “Unemployment\_rate” and “Educational\_special\_needs”. To compare the models, we used the misclassification rate, sensitivity, and specificity.

|  | **Misclassification** | | **Sensitivity** | | **Specificity** | | **AUC** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Training** | **Validation** | **Training** | **Validation** | **Training** | **Validation** | **Training** | **Validation** |
| **Model 1** | 10.96 | 13.14 | 75.54 | 74.98 | 95.22 | 92.53 | - | - |
| **Model 2** | 11.07 | 13.51 | 74.46 | 72.18 | 95.55 | 93.32 | - | - |

As you can see from the results, Models 1 and 2 perform almost identically in both the training and validation sets with slight difference in the values. The misclassification rate is less in model 1 on both training and validation than model2. Sensitivity is higher on the validation set for model1 and higher on training set for model 2 and there is no indication of overfitting since there isn’t a huge gap between the training and validation sets in each model. Specificity is similar on both the models with slight differences. Overall, we decided that Model 1 would be the best model since it has the higher sensitivity on validation set among both models.

**Logistic Regression**

The third method we used for model building is logistic regression. Using logistic regression, we built four different models summarized in the below table.

| **Model** | **Details** |
| --- | --- |
| 1 | No selection method used all variables |
| 2 | Stepwise selection method, used all variables. |
| 3 | Forward selection method, used all variables. |
| 4 | The backward selection method used all variables. |

The Model1 includes all available variables without employing any variable selection method. The purpose is to understand the overall relationship between all variables and the outcome without excluding any. The Model2 employs a stepwise variable selection method, which involves iteratively adding or removing variables based on certain criteria (e.g., p-values, AIC, BIC). Stepwise selection helps identify the most relevant variables while potentially reducing overfitting compared to using all variables.

The Model3 specifically employs a forward selection method, where variables are added to the model one at a time based on certain criteria (e.g., significance level). Forward selection is useful for identifying the most important predictors without considering their interactions. The Model4 specifically employs a backward selection method, where variables are removed from the model one at a time based on certain criteria (e.g., significance level). The Backward selection is useful for simplifying the model by removing non-significant predictors while retaining the most important ones.

|  | **Misclassification** | | **Sensitivity** | | **Specificity** | | **AUC** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Training | **Validation** | **Training** | **Validation** | **Training** | **Validation** | **Training** | **Validation** |
| **Model 1** | 12.82 | 11.76 | 73.91 | 77.49 | 94.27 | 93.96 | 92.5 | 92.2 |
| **Model 2** | 12.82 | 11.76 | 72.26 | 75.65 | 94.35 | 93.80 | 91.9 | 91.4 |
| **Model 3** | 12.82 | 11.76 | 72.26 | 75.65 | 94.35 | 93.80 | 91.9 | 91.4 |
| **Model 4** | 12.4 | 11.4 | 73.83 | 77.12 | 94.23 | 93.64 | 92.4 | 92.04 |

The table presents the performance metrics of four models (Model 1, Model 2, Model 3, and Model 4) in both training and validation datasets. Across all models, the misclassification rates are consistent, ranging from 11.4% to 12.82%. Model 1 demonstrates slightly higher sensitivity in both datasets compared to the other models, indicating its ability to correctly identify positive instances. However, Model 4 exhibits comparable sensitivity values. In terms of specificity, Models 2 and 3 showcase slightly higher values than Models 1 and 4, indicating their effectiveness in correctly identifying negative instances. The area under the curve (AUC) values suggests that Models 1, 2, and 3 have similar discriminatory power, while Model 4 lags slightly behind in the validation dataset.

Overall, Models 1, 2, and 3 exhibit similar performance across most metrics, while Model 4 shows comparable performance but with a slightly lower AUC in the validation dataset.

**Results**

After exploring three different modeling methods, we chose the best models from each to be compared. For CART, the model with important variables: “Curri\_units\_2nd\_sem\_approved”, “tuition\_upToDate”, “Curri\_units\_1st\_sem\_enrolled”, “age\_at\_enrollment”, “curri\_units\_2nd\_sem\_grade”, “course”, “application order”, “Application\_order”, “Curri\_units\_2nd \_sem\_enrolled”, “Curri\_units\_2nd\_sem\_credited”, “Father\_occupation”, “Curri\_units\_1st\_sem\_enrolled”, “ Daytime\_evening\_attendance”, “Curri\_units\_2nd\_sem\_evaluations”, “inflation\_rate”, “Unemployment\_rate” and “Educational\_special\_needs”. For logistic regression, the best model was selected with a backward selection method. The NN model that was chosen was with all the variables from data.

|  | **Misclassification** | | **Sensitivity** | | **Specificity** | | **AUC** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Training** | **Validation** | **Training** | **Validation** | **Training** | **Validation** | **Training** | **Validation** |
| **Model 1(CART)** | 11.43 | 12.93 | 73.04 | 71.33 | 97.89 | 94.63 | 90.36 | 89.25 |
| **Model 2 (NN)** | 10.96 | 13.14 | 75.54 | 74.98 | 95.22 | 92.53 | - | - |
| **Model 3(Logistic)** | 12.4 | 11.4 | 73.83 | 77.12 | 94.23 | 93.64 | 92.4 | 92.04 |

Model 2 has the lowest misclassification rate on the training set but slightly higher on the validation set compared to Model 1 and Model 3. Model 2 has the highest sensitivity on both the training and validation sets, indicating its ability to correctly identify positive cases. Model 2 performs consistently well across all metrics, with relatively high sensitivity on both training and validation sets.

Considering these factors, Model 2 (Neural Network) may be the best choice overall, as it demonstrates strong performance across multiple metrics on both the training and validation sets.

**Discussion and Conclusion**

The goal of our analysis was to determine if we could build a model that would accurately classify a student as a dropout or not. The models we have implemented have about 73% accuracy in finding positive cases and 95% accuracy in finding false cases. The models have shown similar behavior when we have used all variables and when we used only important variables or some selection method.

In conclusion, the logistic regression model (Model 8) outperformed the other models in predicting student dropout based on academic performance. Its superior performance in sensitivity, specificity, and AUC indicates that Model 8 is well-suited for identifying students at risk of dropping out. The simplicity and interpretability of logistic regression make it an attractive choice for this task, especially when considering practical implementation and model transparency. Further the analysis can be extended beyond these parameters as there might be other factors which might influence the students to be a dropout and non dropout.

**Code:**

proc import out=project datafile="/home/u63336679/sasuser.v94/ProjectDataSet.xlsx"

dbms=xlsx replace; sheet = "ProjectDataSet";

run;

data project1;

    set project;

    if Marital\_status in ("1", "2") then Marital\_status = Marital\_status;

    else if Marital\_status in ("3", "4", "5", "6") then Marital\_status = "3";

    if Nationality ="1" then Nationality = Nationality; else Nationality = 0;

    if Application\_mode in ("1", "3", "7", "8", "9") then Application\_mode = "1";

    else if Application\_mode in ("4", "6", "12", "13", "14", "16", "18") then Application\_mode = "2";

    else if Application\_mode in ("2", "5", "10", "11", "15", "17") then Application\_mode = "3";

    if Course in ("1", "4", "7", "8") then Course = "1";

    else if Course in ("2", "5", "15") then Course = "2";

    else if Course in ("6", "12", "13") then Course = "3";

    else if Course in ("9", "3", "10", "11", "14", "17") then Course = "4";

    else if Course in ("16") then Course = "5";

    if Previous\_qualification in ("1", "7", "8", "9", "10", "11", "12", "13") then Previous\_qualification = "1";

    else Previous\_qualification = "2";

    if Mother\_qualification in ("27", "28", "18", "21", "20", "12", "10", "8", "7", "14", "1", "11", "15", "17", "11") then Mother\_qualification = "1";

    else if Mother\_qualification in ("13", "22", "23", "16", "29", "31", "32", "2", "30", "33", "34", "3", "4", "5", "6") then Mother\_qualification = "2";

    else Mother\_qualification = "3";

        if Father\_qualification in ("27", "28", "18", "21", "20", "12", "10", "8", "7", "14", "1", "11", "15", "17", "11") then Father\_qualification = "1";

    else if Father\_qualification in ("13", "22", "23", "16", "29", "31", "32", "2", "30", "33", "34", "3", "4", "5", "6") then Father\_qualification = "2";

    else Father\_qualification = "3";

    if Mother\_occupation in ("2", "3", "5", "9", "11", "14", "15", "16", "17", "18", "19", "20", "21", "22", "27", "28", "29") then Mother\_occupation = "1";

    else if Mother\_occupation in ("4", "7", "8", "6", "23", "24", "25", "26", "36", "37", "38", "39", "40") then Mother\_occupation = "2";

    else Mother\_occupation = "3";

    if Father\_occupation in ("2", "3", "5", "9", "11", "14", "15", "16", "17", "18", "19", "20", "21", "22", "27", "28", "29") then Father\_occupation = "1";

    else if Father\_occupation in ("4", "7", "8", "6", "23", "24", "25", "26", "36", "37", "38", "39", "40") then Father\_occupation = "2";

    else Father\_occupation = "3";

    if Target = "Dropout" then Education\_Status = 1;

    else Education\_Status = 0;

run;

proc freq data=project1;

table Marital\_status Nationality Application\_mode Course Previous\_qualification mother\_qualification Father\_qualification Mother\_occupation Father\_occupation Education\_Status;

Run;

**/\*frequency calculated\*/**

proc freq data=student1;

table 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\_uptodate gender scholarship\_holder international Education\_Status;

run;

/\*Rana\*/

proc import out=Student datafile="/home/u62193061/sasuser.v94/STUDENT1.xlsx"

dbms=xlsx replace;

run;

/\*Correlation \*/

proc corr data=Student;

var Age\_at\_enrollment Curri\_units\_1st\_sem\_credited Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval Curri\_units\_2nd\_sem\_credited Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval Unemployment\_rate Inflation\_rate GDP;

run;

/\*  summary statistics of the macroeconomics data\*/

proc means data=Student n mean median min max std maxdec=3;

var GDP Inflation\_rate Unemployment\_rate;

Run;

/\* academic data at the end of the first semester\*/

proc means data=Student n mean median min max std maxdec=3;

var Curri\_units\_1st\_sem\_credited Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval ;

run;

/\* academic data at the end of the second semester\*/

proc means data=Student n mean median min max std maxdec=3;

var Curri\_units\_2nd\_sem\_credited Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval ;

run;

proc sgplot data=Student;

vbox Age\_at\_enrollment/category=Education\_Status;

run;

proc means data=Student n mean median min max std cv maxdec=3;

var Age\_at\_enrollment;

class Education\_Status;

run;

proc sgplot data=Student;

vbox Unemployment\_rate/category=Education\_Status;

run;

proc sgplot data=Student;

vbox Inflation\_rate/category=Education\_Status;

title "Inflation Rate vs Education Status";

run;

proc sgplot data=Student;

vbox GDP/category=Education\_Status;

run;

proc means data=Student n mean median min max std maxdec=3;

var GDP Inflation\_rate Unemployment\_rate Age\_at\_enrollment;

run;

proc sgplot data=Student;

vbar Gender /stat=pct categoryorder=respasc;;

run;

proc sgplot data=Student;

vbar Education\_Status;

run;

proc freq data=Student;

table Education\_Status;

run;

**/\* PCA\*/**

proc import out=student datafile="/home/u63739299/sasuser.v94/STUDENT1.csv"

dbms=csv replace;

run;

data student2;

set student;

drop 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\_uptodate gender scholarship\_holder international Education\_Status;

run;

proc princomp data=student2;

run;

**/\* CART Model with GINI and partitioning 80% and 20% \*/**

proc hpsplit data=student1 nodes=detail;

    partition fraction(validate=0.2 seed=12345);

    class 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\_uptodate gender scholarship\_holder international Education\_Status;

    model Education\_Status(event="1") = 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\_uptodate

                 gender scholarship\_holder international age\_at\_enrollment Curri\_units\_1st\_sem\_credited

                 Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved

                 Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval Curri\_units\_2nd\_sem\_credited

                 Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved

                 Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval Unemployment\_rate Inflation\_rate GDP;

    grow gini;

    prune cc;

run;

**/\*CART model with variable importance 80% and 20%\*/**

proc hpsplit data=student1 nodes=detail;

    partition fraction(validate=0.2 seed=12345);

    class application\_order course daytime\_evening\_attendance father\_occupation Tuition\_uptodate Education\_Status;

    model Education\_Status(event="1") =  application\_order course daytime\_evening\_attendance father\_occupation

                 Tuition\_uptodate age\_at\_enrollment Curri\_units\_1st\_sem\_credited

                  Curri\_units\_2nd\_sem\_approved Curri\_units\_1st\_sem\_enrolled

                  Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_enrolled Curri\_units\_1st\_sem\_approved

                  Curri\_units\_2nd\_sem\_credited Curri\_units\_2nd\_sem\_evaluations;

    grow gini;

    prune cc;

run;

**/\* Neural Network model and partition 80% and 20% \*/**

proc surveyselect data=student1 method=srs samprate=.2 outall out=student1part3 seed=12345;

run;

proc hpneural data=student1part3;

partition rolevar=selected(train=1);

target Education\_Status/level=nom;

input 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\_uptodate gender scholarship\_holder international/level=nom;

input age\_at\_enrollment Curri\_units\_1st\_sem\_credited Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval Curri\_units\_2nd\_sem\_credited

          Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval Unemployment\_rate Inflation\_rate GDP/level=int;

hidden 34;

train maxiter=1000 numtries=5;

id Education\_Status selected;

score out=student1out3;

run;

**/\*sensitivity and specificity\*/**

proc freq data=student1out3;

table Education\_Status  \* i\_Education\_Status/nocum nocol nopercent;

where selected =1 ;

run;

proc freq data=student1out3;

table Education\_Status \* i\_Education\_Status/nocum nocol nopercent;

where selected =0 ;

run;

**/\*Logistic Regression\*/**

/\* Import file into SAS\*/

proc import out=student datafile="/home/u63739299/sasuser.v94/STUDENT1.csv"

dbms=csv replace;

run;

/\* Creating Dummy Variables \*/

data student3;

set student;

/\* Marital\_status \*/

if Marital\_status = 1 then single= 1 ;else single=0;

if Marital\_status = 2 then married= 1 ;else married=0;

/\* Application\_mode \*/

if Application\_mode = 1 then A\_M1=1; else A\_M1=0;

if Application\_mode = 2 then A\_M2=1; else A\_M2=0;

/\* Course \*/

if Course = 1 then C\_S1=1;else C\_S1=0;

if Course = 2 then C\_S2=1;else C\_S2=0;

if Course = 3 then C\_S3=1;else C\_S3=0;

if Course = 4 then C\_S4=1;else C\_S4=0;

/\* Previous\_qualification \*/

if Previous\_qualification = 1 then P\_Q=1; else P\_Q=0;

/\* Mother\_qualification \*/

if Mother\_qualification = 1 then M\_Q1=1; else M\_Q1=0;

if Mother\_qualification = 2 then M\_Q2=1; else M\_Q2=0;

/\* Father\_qualification \*/

if Father\_qualification = 1 then F\_Q1=1; else F\_Q1=0;

if Father\_qualification = 2 then F\_Q2=1; else F\_Q2=0;

/\* Mother\_occupation \*/

if Mother\_occupation = 1 then M\_O1=1; else M\_O1=0;

if Mother\_occupation = 2 then M\_O2=1; else M\_O2=0;

/\* Father\_occupation \*/

if Father\_occupation = 1 then F\_O1=1; else F\_O1=0;

if Father\_occupation = 2 then F\_O2=1; else F\_O2=0;

run;

/\* Data Partition \*/

proc surveyselect data=student3 method=srs samprate=.8 outall out=student3part seed=12345;

run;

data student3train student3valid;

set student3part;

if selected=1 then output student3train; else output student3valid;

run;

/\* Logistic Model using all Variables \*/

proc logistic data=student3train outmodel=student3model;

model Education\_Status(event="1")= single married A\_M1 A\_M2 C\_S1 C\_S2 C\_S3 C\_S4 application\_order

daytime\_evening\_attendance P\_Q Nationality M\_Q1 M\_Q2 F\_Q1 F\_Q2

M\_O1 M\_O2 F\_O1 F\_O2 Displaced Educational\_special\_needs debtor Tuition\_uptodate

gender scholarship\_holder international age\_at\_enrollment Curri\_units\_1st\_sem\_credited

Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved

Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval Curri\_units\_2nd\_sem\_credited

Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved

Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval Unemployment\_rate Inflation\_rate GDP;

run;

proc logistic inmodel=student3model;

score data=student3train fitstat out=student3trainout;

score data=student3valid fitstat out=student3validout;

run;

proc freq data=student3trainout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;

proc freq data=student3validout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;

/\* Logistic Model using stepwise selection \*/

proc logistic data=student3train outmodel=student3model;

model Education\_Status(event="1")= single married A\_M1 A\_M2 C\_S1 C\_S2 C\_S3 C\_S4 application\_order

daytime\_evening\_attendance P\_Q Nationality M\_Q1 M\_Q2 F\_Q1 F\_Q2

M\_O1 M\_O2 F\_O1 F\_O2 Displaced Educational\_special\_needs debtor Tuition\_uptodate

gender scholarship\_holder international age\_at\_enrollment Curri\_units\_1st\_sem\_credited

Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved

Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval Curri\_units\_2nd\_sem\_credited

Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved

Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval Unemployment\_rate Inflation\_rate GDP/selection=stepwise;

run;

proc logistic inmodel=student3model;

score data=student3train fitstat out=student3trainout;

score data=student3valid fitstat out=student3validout;

run;

proc freq data=student3trainout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;

proc freq data=student3validout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;

/\* Logistic Model using forward selection \*/

proc logistic data=student3train outmodel=student3model;

model Education\_Status(event="1")= single married A\_M1 A\_M2 C\_S1 C\_S2 C\_S3 C\_S4 application\_order

daytime\_evening\_attendance P\_Q Nationality M\_Q1 M\_Q2 F\_Q1 F\_Q2

M\_O1 M\_O2 F\_O1 F\_O2 Displaced Educational\_special\_needs debtor Tuition\_uptodate

gender scholarship\_holder international age\_at\_enrollment Curri\_units\_1st\_sem\_credited

Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved

Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval Curri\_units\_2nd\_sem\_credited

Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved

Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval Unemployment\_rate Inflation\_rate GDP/selection=forward;

run;

proc logistic inmodel=student3model;

score data=student3train fitstat out=student3trainout;

score data=student3valid fitstat out=student3validout;

run;

proc freq data=student3trainout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;

proc freq data=student3validout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;

/\* Logistic Model using backward selection \*/

proc logistic data=student3train outmodel=student3model;

model Education\_Status(event="1")= single married A\_M1 A\_M2 C\_S1 C\_S2 C\_S3 C\_S4 application\_order

daytime\_evening\_attendance P\_Q Nationality M\_Q1 M\_Q2 F\_Q1 F\_Q2

M\_O1 M\_O2 F\_O1 F\_O2 Displaced Educational\_special\_needs debtor Tuition\_uptodate

gender scholarship\_holder international age\_at\_enrollment Curri\_units\_1st\_sem\_credited

Curri\_units\_1st\_sem\_enrolled Curri\_units\_1st\_sem\_evaluations Curri\_units\_1st\_sem\_approved

Curri\_units\_1st\_sem\_grade Curri\_units\_1st\_sem\_woeval Curri\_units\_2nd\_sem\_credited

Curri\_units\_2nd\_sem\_enrolled Curri\_units\_2nd\_sem\_evaluations Curri\_units\_2nd\_sem\_approved

Curri\_units\_2nd\_sem\_grade Curri\_units\_2nd\_sem\_woeval Unemployment\_rate Inflation\_rate GDP/selection=backward;

run;

proc logistic inmodel=student3model;

score data=student3train fitstat out=student3trainout;

score data=student3valid fitstat out=student3validout;

run;

proc freq data=student3trainout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;

proc freq data=student3validout;

table f\_Education\_Status\*i\_Education\_Status/nocol nopercent nocum;

run;