Database

Data of Academic Performance evolution for Engineering Students

This data article presents data on the results in national assessments for secondary and university education in engineering students. The data contains academic, social, economic information for 12,411 students. The data were obtained by orderly crossing the databases of the Colombian Institute for the

https://data.mendeley.com/datasets/83tcx8psxv/1

https://www.sciencedirect.com/science/article/pii/S2352340920304315?via=ihub

Description :

- Data represents the national assessments results for secondary and university education for students taking an engineering course
- Data considers academic, social and economic information of students
- Database was obtained orderly crossing databases of the Colombian Institute for Evaluation of Engineering
- Gender distribution 5043 (40.63%) for women and 7368 (59.37%) for men

How the database was acquired :

- Collection, adaptation and adjustment of information was carried out by the Colombian Institute for the Evaluation of Education.

Value of the data :

- The features are fit to make predictions and classification models of academic, social and economic features.

- The data was gathered for both High School and University students.
- The database uses ID’s to guarantee the anonymity of students.

Data Description :

- Database contains 12411 observations and 44 variables.
- The database consists of personal information (categorical) and assessments (numerical).

Features present in the database :

▼ Code

```
import numpy as np
import pandas as pd
import sklearn.linear_model, sklearn.datasets
from sklearn.preprocessing import StandardScaler, MinMaxScale
import matplotlib.pyplot as plt
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error, r2_score, mea
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer

testData = pd.read_csv('Database/DatabaseTest.csv')

pd.set_option('display.max_columns', None)
print(testData)
```

	UNIVERSITY	ACADEMIC_PROGRAM	QR_PRO	CR_PRO	CC_PRO	ENG_PRO	WC_PRO	FEP_PRO	G_SC	PERCENTILE	2ND_DECILE	QUARTILE	SEL	SEL_IHE
0	UNIVERSIDAD DE SANTANDER – UDES–BUCARAMANGA	INDUSTRIAL ENGINEERING	71	93	71	93	79	181	180	91	5	4	2	2
1	UNIVERSIDAD DE LOS ANDES–BOGOTÁ D.C.	INDUSTRIAL ENGINEERING	97	38	86	98	78	201	182	92	5	4	4	4
2	UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA UNA...	ELECTRONIC ENGINEERING	17	1	18	43	22	113	113	7	1	1	1	1
3	UNIVERSIDAD CATOLICA DE PEREIRA–PEREIRA	INDUSTRIAL ENGINEERING	65	35	76	80	48	137	157	67	4	3	2	2
4	UNIVERSIDAD INDUSTRIAL DE SANTANDER–BUCARAMANGA	INDUSTRIAL ENGINEERING	94	94	98	100	71	189	198	98	5	4	4	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
12406	UNIVERSIDAD ECCI–BOGOTÁ D.C.	MECHATRONICS ENGINEERING	88	71	86	87	65	142	176	88	5	4	2	2
12407	INSTITUCION UNIVERSITARIA DE COLOMBIA – UNIVER...	INDUSTRIAL ENGINEERING	46	39	44	11	0	127	107	4	1	1	4	2
12408	UNIVERSIDAD TECNOLOGICA DE BOLIVAR–CARTAGENA	INDUSTRIAL ENGINEERING	98	88	90	81	87	192	188	95	5	4	2	2
12409	UNIVERSIDAD TECNOLOGICA DE BOLIVAR–CARTAGENA	CIVIL ENGINEERING	60	80	51	8	42	121	146	50	3	3	3	2
12410	UNIVERSIDAD DE LOS ANDES–BOGOTÁ D.C.	INDUSTRIAL ENGINEERING	83	95	91	79	47	193	178	89	5	4	2	4

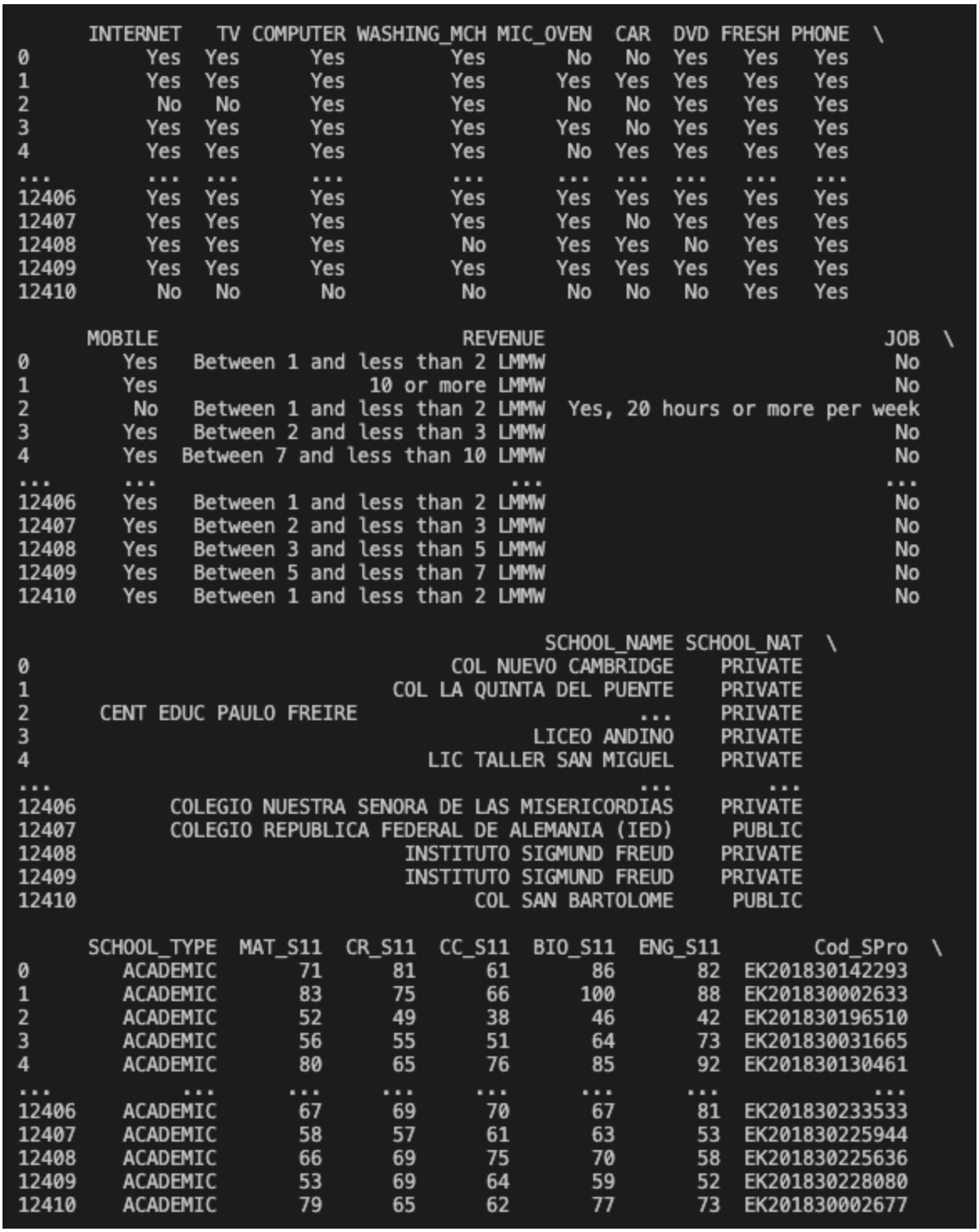
[12411 rows x 45 columns]

```
(base) cristiandumbravanu@Cristians-MBP My Work % /usr/local/bin/python3 "/Users/cristiandumbravanu/Desktop/Final Year Project/My Work/Linear Regression/Linear Regression Main.py"
              COD_S11 GENDER                          EDU_FATHER  \
0      SB11201210000129      F  Incomplete Professional Education
1      SB11201210000137      F                 Complete Secundary
2      SB11201210005154      M                          Not sure
3      SB11201210007504      F                          Not sure
4      SB11201210007548      M    Complete professional education
...                 ...    ...                             ...
12406  SB11201420568705      M                           Ninguno
12407  SB11201420573045      M    Complete professional education
12408  SB11201420578809      M   Complete technique or technology
12409  SB11201420578812      F    Complete professional education
12410  SB11201420583232      M                 Complete Secundary

                          EDU_MOTHER  \
0      Complete technique or technology
1       Complete professional education
2                             Not sure
3                             Not sure
4       Complete professional education
...                                ...
12406               Complete Secundary
12407               Complete Secundary
12408  Complete technique or technology
12409   Complete professional education
12410                Complete primary

                                  OCC_FATHER                  OCC_MOTHER  \
0      Technical or professional level employee                        Home
1                                Entrepreneur      Independent professional
2                                 Independent                        Home
3                            Other occupation                 Independent
4                                   Executive                        Home
...                                       ...                         ...
12406                        Other occupation  Auxiliary or Administrative
12407                               Executive             Other occupation
12408                                 Retired                        Home
12409                 Independent professional          Small entrepreneur
12410                               Independent                        Home

        STRATUM                          SISBEN PEOPLE_HOUSE  Unnamed: 9  \
0      Stratum 4  It is not classified by the SISBEN        Three         NaN
1      Stratum 5  It is not classified by the SISBEN        Three         NaN
2      Stratum 2                         Level 2         Five         NaN
3      Stratum 2  It is not classified by the SISBEN        Three         NaN
4      Stratum 4  It is not classified by the SISBEN          One         NaN
...          ...                            ...          ...         ...
12406  Stratum 2  It is not classified by the SISBEN          Six         NaN
12407  Stratum 2                         Level 2         Five         NaN
12408  Stratum 2                         Level 2         Five         NaN
12409  Stratum 3  It is not classified by the SISBEN        Seven         NaN
12410  Stratum 3                         Level 1         Four         NaN
```



	INTERNET	TV	COMPUTER	WASHING_MCH	MIC_OVEN	CAR	DVD	FRESH	PHONE
0	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
2	No	No	Yes	Yes	No	No	Yes	Yes	Yes
3	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
4	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
...	...	...	...	...	...	...	...	...	...
12406	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
12407	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
12408	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes
12409	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
12410	No	No	No	No	No	No	No	Yes	Yes

	MOBILE	REVENUE	JOB
0	Yes	Between 1 and less than 2 LMMW	No
1	Yes	10 or more LMMW	No
2	No	Between 1 and less than 2 LMMW	Yes, 20 hours or more per week
3	Yes	Between 2 and less than 3 LMMW	No
4	Yes	Between 7 and less than 10 LMMW	No
...	...	...	...
12406	Yes	Between 1 and less than 2 LMMW	No
12407	Yes	Between 2 and less than 3 LMMW	No
12408	Yes	Between 3 and less than 5 LMMW	No
12409	Yes	Between 5 and less than 7 LMMW	No
12410	Yes	Between 1 and less than 2 LMMW	No

	SCHOOL_NAME	SCHOOL_NAT
0	COL NUEVO CAMBRIDGE	PRIVATE
1	COL LA QUINTA DEL PUENTE	PRIVATE
2	CENT EDUC PAULO FREIRE	PRIVATE
3	LICEO ANDINO	PRIVATE
4	LIC TALLER SAN MIGUEL	PRIVATE
...	...	...
12406	COLEGIO NUESTRA SENORA DE LAS MISERICORDIAS	PRIVATE
12407	COLEGIO REPUBLICA FEDERAL DE ALEMANIA (IED)	PUBLIC
12408	INSTITUTO SIGMUND FREUD	PRIVATE
12409	INSTITUTO SIGMUND FREUD	PRIVATE
12410	COL SAN BARTOLOME	PUBLIC

	SCHOOL_TYPE	MAT_S11	CR_S11	CC_S11	BIO_S11	ENG_S11	Cod_SPro
0	ACADEMIC	71	81	61	86	82	EK201830142293
1	ACADEMIC	83	75	66	100	88	EK201830002633
2	ACADEMIC	52	49	38	46	42	EK201830196510
3	ACADEMIC	56	55	51	64	73	EK201830031665
4	ACADEMIC	80	65	76	85	92	EK201830130461
...	...	...	...	...	...	...	...
12406	ACADEMIC	67	69	70	67	81	EK201830233533
12407	ACADEMIC	58	57	61	63	53	EK201830225944
12408	ACADEMIC	66	69	75	70	58	EK201830225636
12409	ACADEMIC	53	69	64	59	52	EK201830228080
12410	ACADEMIC	79	65	62	77	73	EK201830002677

Features :

The academic assessment is recorded at two moments of the student life :

1. Scores of the national standardised test at the final year of high school :

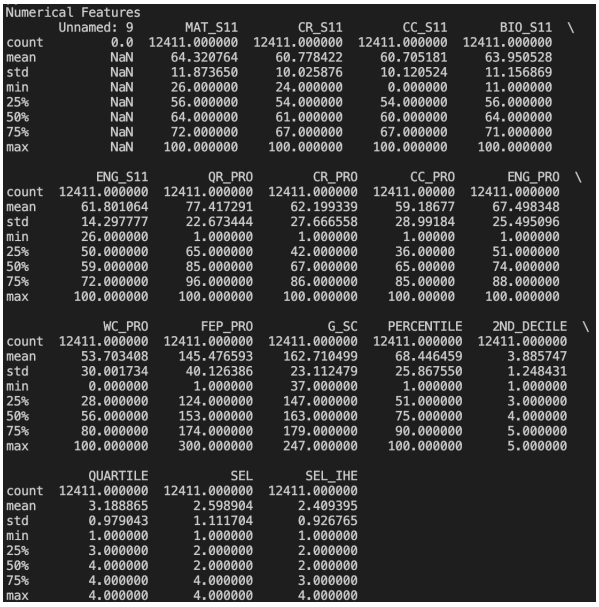
- Saber 11 (S11) : Score of the national test at the final year of high school, evaluating five generic academic competences :
    - Maths (MAT_S11) : assess the skills of facing problems with the use of math tools
    - Critical Reading (CR_11) : assesses the skills of understanding, interpreting and evaluating texts
    - Citizen Competences (CC_S11) : assess student's skills that allow him to understand the social world from the point of view of social sciences and the role of a citizen
    - Biology (BIO_S11) : assesses how a student explains the phenomena of nature occurence with the employment of observations, patterns and scientific concepts
    - English (ENG_S11) : asses the competence of communication in English

2. Second moment of academic assessment is the final year of the Professional Career on Engineering (University) also recorded on the national standardised test for higher education using SABER PRO. Similarly to the Saber 11 tests, it uses five academic measurements :

- Critical Reading (CR_SPRO) : assess the ability to understand text and the critical approach to it
- Quantitative Reasoning (QR_PRO) : assess the ability to understand and manipulate quantitative data in forms of diagrams, graphs or tables
- Citizen competence (CC_PRO) : assess the concept of citizenship and the existence within the Colombian constitution
- Written communication (WC_PRO) : assess the ability to transform ideas into writing based on a topic
- English (ENG_PRO) : asses the competence of communicating using English
- Sisben : the economic aid that the Columbian government gives to low income families to offer them an improved quality of life

---

Feature stats :

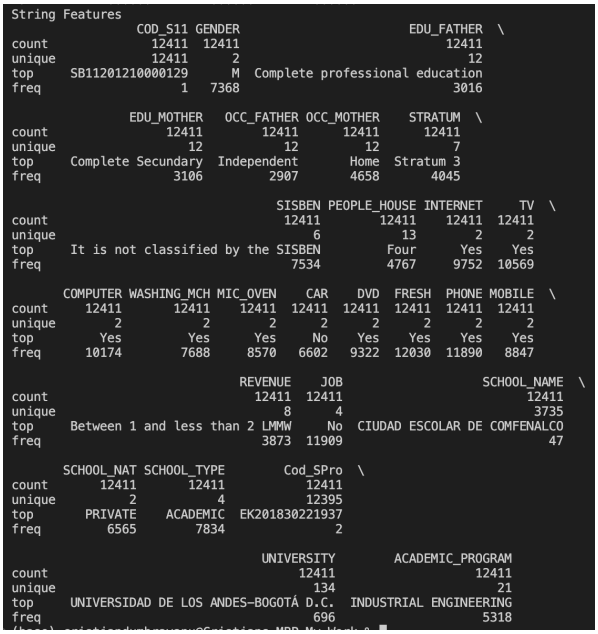


```
Numerical Features
       Unnamed: 9       MAT_S11       CR_S11       CC_S11       BIO_S11  \
count         0.0  12411.000000  12411.000000  12411.000000  12411.000000
mean          NaN     64.320764     60.778422     60.705181     63.950528
std           NaN     11.873650     10.025876     10.120524     11.156869
min           NaN     26.000000     24.000000      0.000000     11.000000
25%           NaN     56.000000     54.000000     54.000000     56.000000
50%           NaN     64.000000     61.000000     60.000000     64.000000
75%           NaN     72.000000     67.000000     67.000000     71.000000
max           NaN    100.000000    100.000000    100.000000    100.000000

            ENG_S11       QR_PRO       CR_PRO       CC_PRO       ENG_PRO  \
count  12411.000000  12411.000000  12411.000000  12411.00000  12411.000000
mean      61.801064     77.417291     62.199339     59.18677     67.498348
std       14.297777     22.673444     27.666558     28.99184     25.495096
min       26.000000      1.000000      1.000000      1.00000      1.000000
25%       50.000000     65.000000     42.000000     36.00000     51.000000
50%       59.000000     85.000000     67.000000     65.00000     74.000000
75%       72.000000     96.000000     86.000000     85.00000     88.000000
max      100.000000    100.000000    100.000000    100.00000    100.000000

             WC_PRO       FEP_PRO         G_SC    PERCENTILE    2ND_DECILE  \
count  12411.000000  12411.000000  12411.000000  12411.000000  12411.000000
mean      53.703408    145.476593    162.710499     68.446459      3.885747
std       30.001734     40.126386     23.112479     25.867550      1.248431
min        0.000000      1.000000     37.000000      1.000000      1.000000
25%       28.000000    124.000000    147.000000     51.000000      3.000000
50%       56.000000    153.000000    163.000000     75.000000      4.000000
75%       80.000000    174.000000    179.000000     90.000000      5.000000
max      100.000000    300.000000    247.000000    100.000000      5.000000

          QUARTILE           SEL      SEL_IHE
count  12411.000000  12411.000000  12411.000000
mean       3.188865      2.598904      2.409395
std        0.979043      1.111704      0.926765
min        1.000000      1.000000      1.000000
25%        3.000000      2.000000      2.000000
50%        4.000000      2.000000      2.000000
75%        4.000000      4.000000      3.000000
max        4.000000      4.000000      4.000000
```



```
String Features
             COD_S11 GENDER                          EDU_FATHER  \
count            12411  12411                               12411
unique           12411      2                                  12
top     SB11201210000129      M  Complete professional education
freq                 1   7368                                3016

            EDU_MOTHER   OCC_FATHER OCC_MOTHER    STRATUM  \
count            12411        12411      12411      12411
unique              12           12         12          7
top     Complete Secundary  Independent       Home  Stratum 3
freq                  3106         2907       4658       4045

                                  SISBEN PEOPLE_HOUSE INTERNET     TV  \
count                              12411        12411    12411  12411
unique                                 6           13        2      2
top     It is not classified by the SISBEN         Four      Yes    Yes
freq                                7534         4767     9752  10569

       COMPUTER WASHING_MCH MIC_OVEN    CAR    DVD  FRESH  PHONE MOBILE  \
count     12411       12411    12411  12411  12411  12411  12411  12411
unique        2           2        2      2      2      2      2      2
top         Yes         Yes      Yes     No    Yes    Yes    Yes    Yes
freq      10174        7688     8570   6602   9322  12030  11890   8847

                                REVENUE    JOB                    SCHOOL_NAME  \
count                             12411  12411                          12411
unique                                8      4                           3735
top     Between 1 and less than 2 LMMW     No  CIUDAD ESCOLAR DE COMFENALCO
freq                               3873  11909                             47

       SCHOOL_NAT SCHOOL_TYPE        Cod_SPro  \
count       12411       12411           12411
unique          2           4           12395
top       PRIVATE    ACADEMIC  EK201830221937
freq         6565        7834               2

                                  UNIVERSITY        ACADEMIC_PROGRAM
count                                  12411                   12411
unique                                   134                      21
top     UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.  INDUSTRIAL ENGINEERING
freq                                    696                    5318
```

Database preprocessing :

- I have added two new columns of Average Score and Pass_Fail

The new Database stats :

▼ Code

```
import numpy as np
import pandas as pd
import sklearn.linear_model
from sklearn.preprocessing import StandardScaler, MinMaxScale
import matplotlib.pyplot as plt
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import confusion_matrix, accuracy_score,
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer

dataBase = pd.read_csv('Database/DatabaseTest.csv')

pd.set_option('display.max_columns', None)
```

```
# Calculate the average of SABER PRO'S 'QR_PRO', 'CR_PRO', 'C
dataBase["Average_Score_SABER_PRO"] = dataBase[['QR_PRO', 'CR
dataBase["Pass_Fail_SABER_PRO"] = np.where(dataBase['Average_

# Calculate the average of Saber 11 competences
dataBase["Average_Score_Saber11"] = dataBase[['MAT_S11','CR_S
dataBase["Pass_Fail_Saber11"] = np.where(dataBase['Average_Sc

pass_fail_counts_SABER_PRO = dataBase['Pass_Fail_SABER_PRO']
pass_fail_counts_Saber11 = dataBase['Pass_Fail_Saber11'].valu
print(pass_fail_counts_SABER_PRO)
print(pass_fail_counts_Saber11)

print("Categorical Features")
print(dataBase.select_dtypes(exclude=np.number).describe())
print("Numerical Features")
print(dataBase.select_dtypes(include=np.number).describe())
```

```
Pass_Fail_SABER_PRO
Pass    7537
Fail    4874
Name: count, dtype: int64
Pass_Fail_Saber11
Pass    7135
Fail    5276
Name: count, dtype: int64
Categorical Features
                COD_S11 GENDER                          EDU_FA
count             12411  12411                               12
unique            12411      2
top     SB11201210000129      M  Complete professional educat
freq                  1   7368                               3

               EDU_MOTHER   OCC_FATHER OCC_MOTHER    STRATU
```

```
count                12411        12411      12411      1241
unique                  12           12         12         
top     Complete Secundary  Independent       Home  Stratum 
freq                  3106         2907       4658       404

                                    SISBEN PEOPLE_HOUSE INTE
count                                12411        12411    12
unique                                   6           13      
top     It is not classified by the SISBEN         Four      
freq                                  7534         4767     9

       COMPUTER WASHING_MCH MIC_OVEN    CAR    DVD  FRESH  PH
count     12411       12411    12411  12411  12411  12411  12
unique        2           2        2      2      2      2     
top         Yes         Yes      Yes     No    Yes    Yes     
freq      10174        7688     8570   6602   9322  12030  11

                               REVENUE    JOB                
count                            12411  12411                
unique                               8      4                
top     Between 1 and less than 2 LMMW     No  CIUDAD ESCOLAR
freq                              3873  11909                

       SCHOOL_NAT SCHOOL_TYPE        Cod_SPro  \               
count       12411       12411           12411                
unique          2           4           12395                
top       PRIVATE    ACADEMIC  EK201830221937                
freq         6565        7834               2                

                                    UNIVERSITY        ACADEMIC_
count                                    12411                 
unique                                     134                 
top     UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.  INDUSTRIAL ENGI
freq                                       696                 

       Pass_Fail_SABER_PRO Pass_Fail_Saber11
```

```
count                12411             12411
unique                   2                 2
top                   Pass              Pass
freq                  7537              7135
Numerical Features
       Unnamed: 9       MAT_S11        CR_S11        CC_S11
count         0.0  12411.000000  12411.000000  12411.000000
mean          NaN     64.320764     60.778422     60.705181
std           NaN     11.873650     10.025876     10.120524
min           NaN     26.000000     24.000000      0.000000
25%           NaN     56.000000     54.000000     54.000000
50%           NaN     64.000000     61.000000     60.000000
75%           NaN     72.000000     67.000000     67.000000
max           NaN    100.000000    100.000000    100.000000

            ENG_S11        QR_PRO        CR_PRO       CC_PRO
count  12411.000000  12411.000000  12411.000000  12411.00000
mean      61.801064     77.417291     62.199339     59.18677
std       14.297777     22.673444     27.666558     28.99184
min       26.000000      1.000000      1.000000      1.00000
25%       50.000000     65.000000     42.000000     36.00000
50%       59.000000     85.000000     67.000000     65.00000
75%       72.000000     96.000000     86.000000     85.00000
max      100.000000    100.000000    100.000000    100.00000

             WC_PRO       FEP_PRO          G_SC    PERCENTILE
count  12411.000000  12411.000000  12411.000000  12411.000000
mean      53.703408    145.476593    162.710499     68.446459
std       30.001734     40.126386     23.112479     25.867550
min        0.000000      1.000000     37.000000      1.000000
25%       28.000000    124.000000    147.000000     51.000000
50%       56.000000    153.000000    163.000000     75.000000
75%       80.000000    174.000000    179.000000     90.000000
max      100.000000    300.000000    247.000000    100.000000

           QUARTILE           SEL       SEL_IHE  Average_Scor
```

```
count  12411.000000  12411.000000  12411.000000              1
mean       3.188865      2.598904      2.409395
std        0.979043      1.111704      0.926765
min        1.000000      1.000000      1.000000
25%        3.000000      2.000000      2.000000
50%        4.000000      2.000000      2.000000
75%        4.000000      4.000000      3.000000
max        4.000000      4.000000      4.000000

       Average_Score_Saber11
count           12411.000000
mean               62.311192
std                 9.642924
min                35.800000
25%                55.200000
50%                61.800000
75%                69.000000
max                95.600000
```

```
testData['Average_Score'] = testData[['QR_PRO', 'CR_PRO', 'CC_PR
```

- The column Average Score was made using the already existing features of :
  - QR_PRO / Quantitative Reasoning : assessment of the ability to understand and manipulate quantitative data.
  - CR_PRO / Critical Reading : assessment of the ability to understand a text and be able to critically approach it.
  - CC_PRO / Citizen Competence : refers to the ability, knowledge and attitude a student possess to effectively participate in society as a responsible and engaged citizen.
  - ENG_PRO / English : assessment of the competence to communicate effectively using English
  - WC_PRO / Written Communication : assessment of student's ability to write ideas of a related topic.

```
testData['Pass_Fail'] = np.where(testData['Average_Score'] >= 6(
```

- The column of Pass_Fail is dependent to the Average Score of SABER PRO's mean.
  - If the mean is Below 60 it's a fail, if the mean is equal or over 60 its a pass.

```
dataBase["Average_Score_Saber11"] = dataBase[['MAT_S11','CR_S11
```

- The column of Average Score of Sober 11 refers to the mean value of the features :
  - MAT_S11 / Mathematics : assessment of how students solve problems with the use of maths
  - CR_S11 / Critical Reading : assessment of how students understand, interpret and evaluate texts.
  - CC_S11 / Citizen Competences : assessment of student's knowledge and skills that allow him to understand the social world with the use of social sciences to fulfil the role of a citizen.
  - BIO_S11 / Biology : assessment of the student's ability to explain natural phenomenas based on observations, patterns and scientifical concepts.
  - ENG_S11 / English : assessment of the competence to communicate in English

```
dataBase["Pass_Fail_Saber11"] = np.where(dataBase['Average_Score
```

- This column uses the mean values of the Average_Score_Saber11 and using the threshold of 60 determines if the student has passed or failed.

The New Column Stats :

```
Pass_Fail_SABER_PRO
Pass    7537
Fail    4874
```

```
Name: count, dtype: int64
Pass_Fail_Saber11
Pass    7135
Fail    5276
Name: count, dtype: int64

Pass_Fail_SABER_PRO        Pass_Fail_Saber11
count                12411             12411
unique                   2                 2
top                   Pass              Pass
freq                  7537              7135

Average_Score_Saber11
count           12411.000000
mean               62.311192
std                 9.642924
min                35.800000
25%                55.200000
50%                61.800000
75%                69.000000
max                95.600000
```

And :