Building and Analyzing Binary Classification models of students success

RECOMMENDATIONS FOR UNIVERSITY OFFICIALS: HELPING YOUR FAILING STUDENTS TO SUCCEED

FLATIRON SCHOOL

Phase 3 Final Project

Authors: Elena Kazakova

Cohort: DS02222021

Instructor: James Irving

Summary

Analysis of engineering students performance

- **Colombia**
- Gathered and compiled between **2012** and **2018**

The resulting models provide insight into

- **What factors** affect engineering students performance
- **What Programs/Universities** are the most successful

Outline of the presentation

- Business Problem
- Data
- Final Model
- Results
- Conclusions

Business Problem

- Build a predictive model(s) of graduating Colombian Engineering Students to help

  University officials to identify failing students early

- Build ranking lists of the best and the worst Programs/Universities to help

  Prospective students to choose a program

Data

- Colombia engineering students' test scores and socio-economic characteristics
- Between **2012** and **2018**
- **21411** records
- **44** categorical and numerical variables
- The column names are listed below (see Appendix for detailed description)

GENDER
EDU_FATHER
EDU_MOTHER
OCC_FATHER
OCC_MOTHER
STRATUM
SISBEN
PEOPLE_HOUSE
SEL
SEL_IHE

INTERNET
TV
COMPUTER
WASHING_MCH
MIC_OVEN
CAR
DVD
FRESH
PHONE
MOBILE

REVENUE
JOB
SCHOOL_NAME
SCHOOL_NAT
SCHOOL_TYPE
UNIVERSITY
ACADEMIC_PROGRAM
PERCENTILE
2ND_DECILE

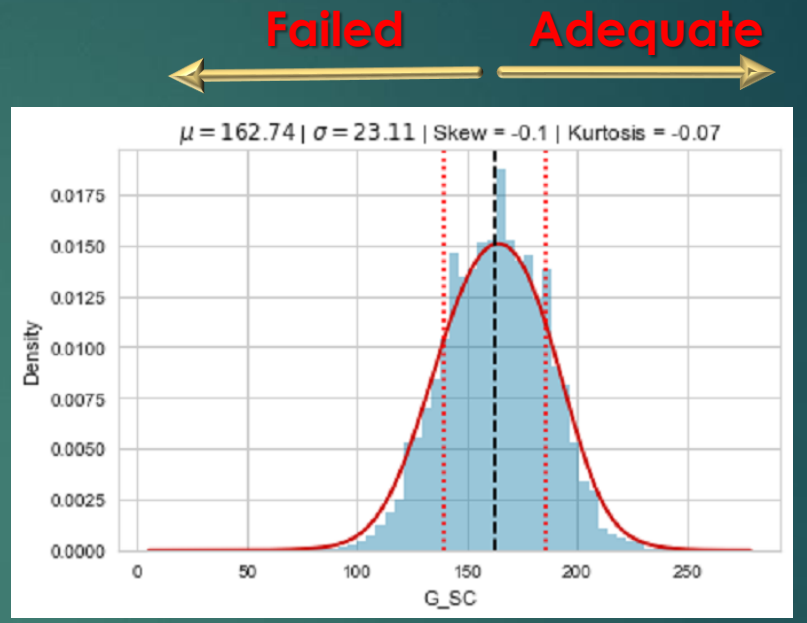
QR_PRO
CR_PRO
CC_PRO
ENG_PRO
WC_PRO
FEP_PRO
G_SC

MAT_S11
CR_S11
CC_S11
BIO_S11
ENG_S11

Modifications to the Data

Adjusted target to a binary format:

(1,0): Failed vs. Adequate



Removed features:

- **Test IDs**
- **Professional subject test scores**
- **Program and University**

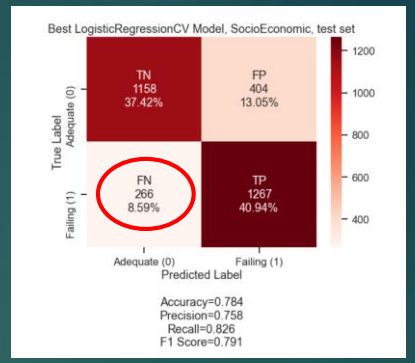
Clean-up:

- **Translation**
- **Replacing "Unknown" value**
- **Removing records with errors**

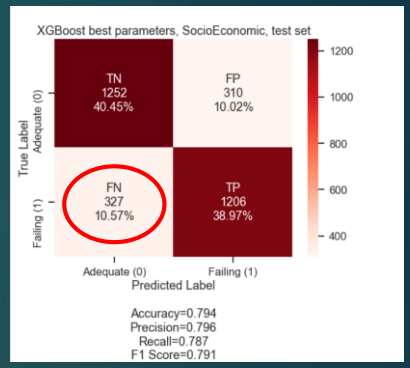
Final Models

- **Sixteen** models tested
- Separate models
  - **Two** with Socio-economic/test scores
  - **Two** with University/Program features
- **Two** front runners **for each case** to validate the results
- Accuracy/Recall rates:
  - *How many did we get right where it matters*

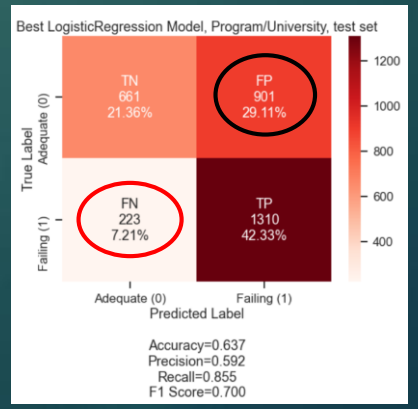
Model #1



Model #2

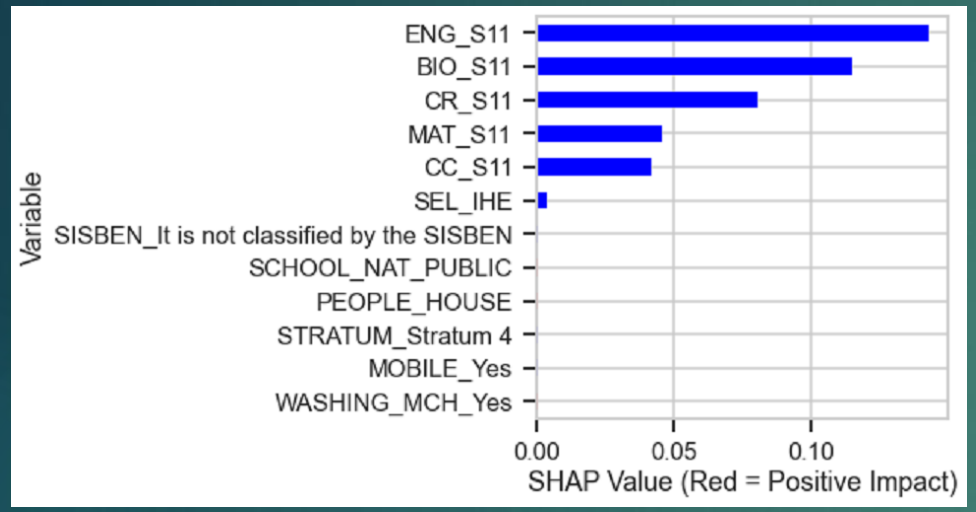


Model #3



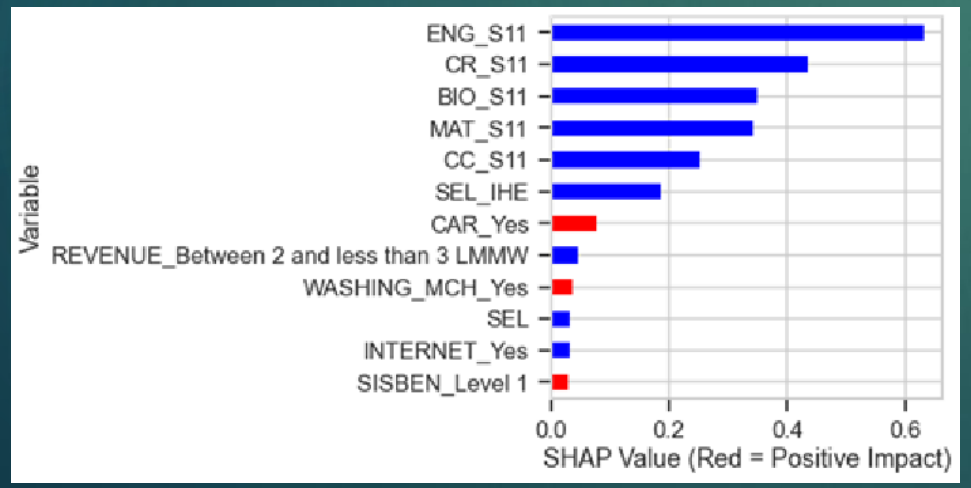
Final Models (continued)

Model #1



- ENG_S11
- ***BIO_S11***
- ***CR_S11***
- MAT_S11
- CC_S11
- SEL_IHE

Model #2

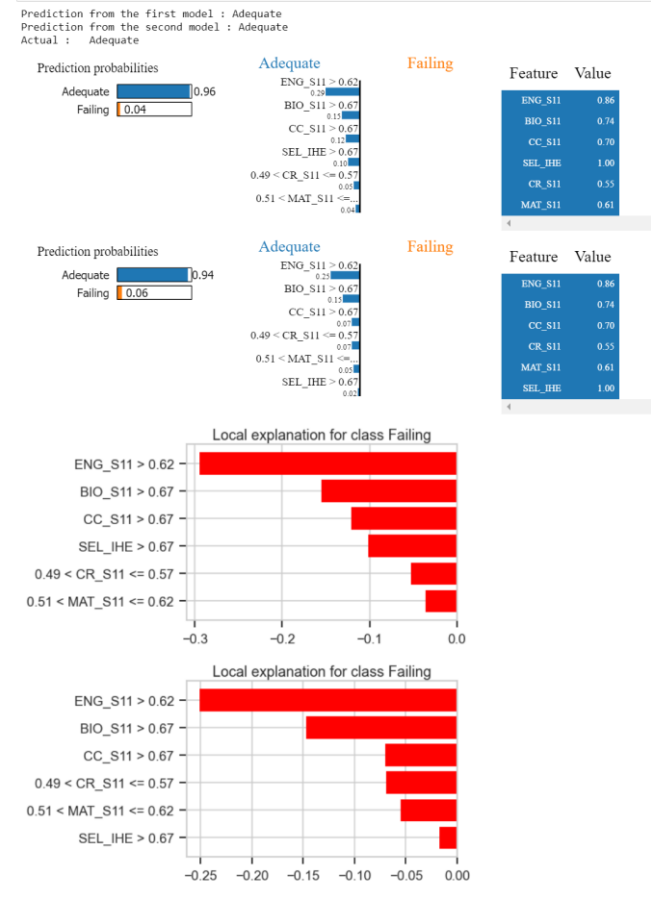


- ENG_S11
- ***CR_S11***
- ***BIO_S11***
- MAT_S11
- CC_S11
- SEL_IHE

- **Agreement** between models
- **HS scores** more important than **socio-economic school level**

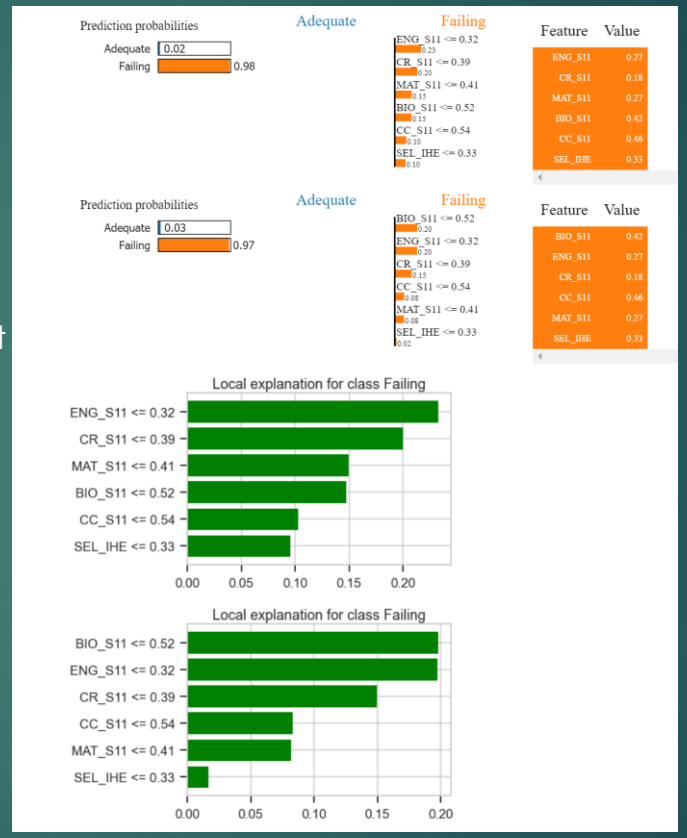
Use cases

#1



- **Outstanding** HS test results
- In a program at a university with a **wealthy** student population
- Both models predictions **in agreement** with the actual result
- Performance: **Adequate**

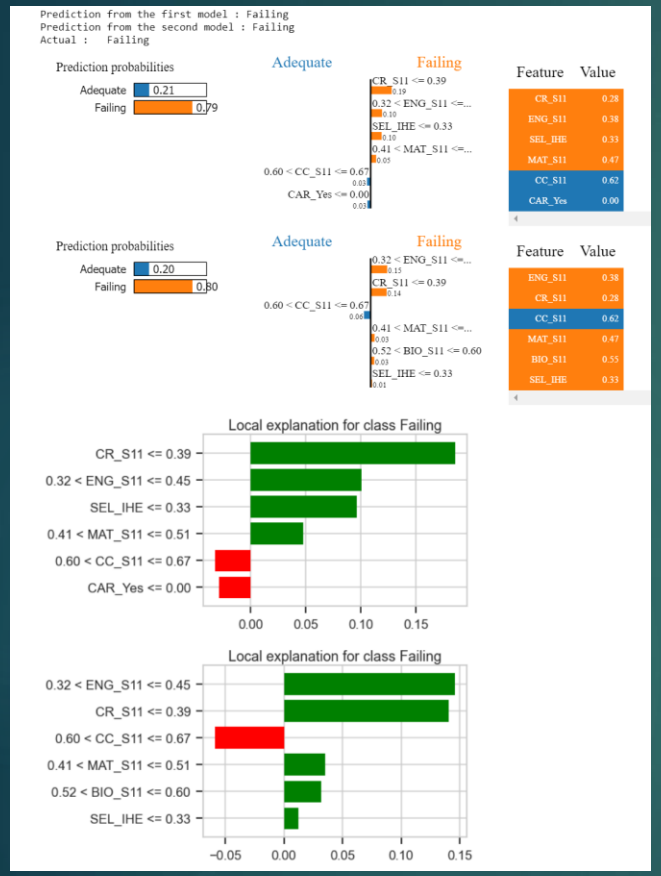
#2



- **Poor** HS test results
- In a program at a university with a **low-income** student population
- Both models predictions **in agreement** with the actual result
- Performance: **Failed**

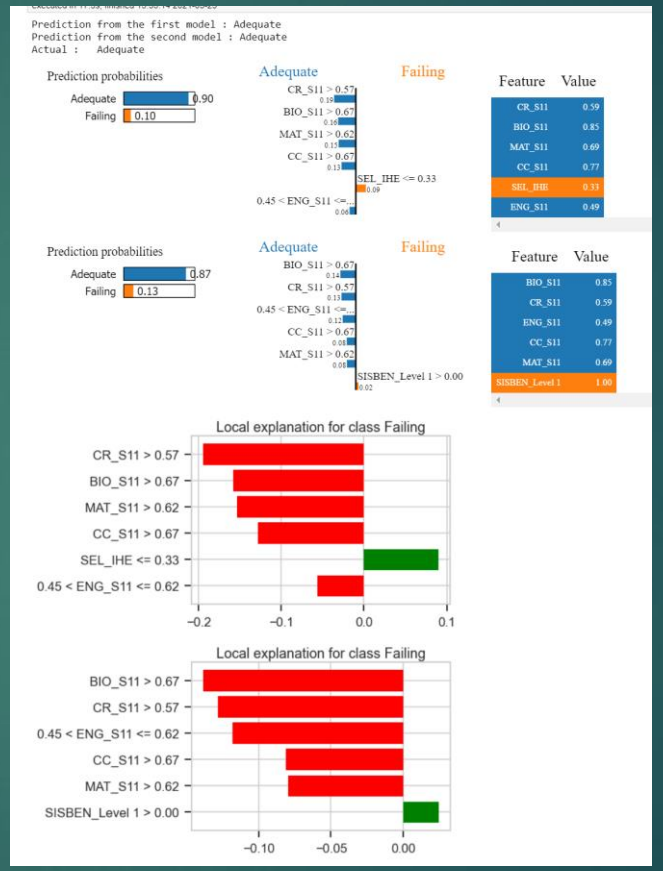
Use cases (continued)

#3



- **Poor** HS test results (except Civics)
- In a program at a university with a **low-income** student population
- Both models predictions **in agreement** with the actual result
- Performance: **Failed**

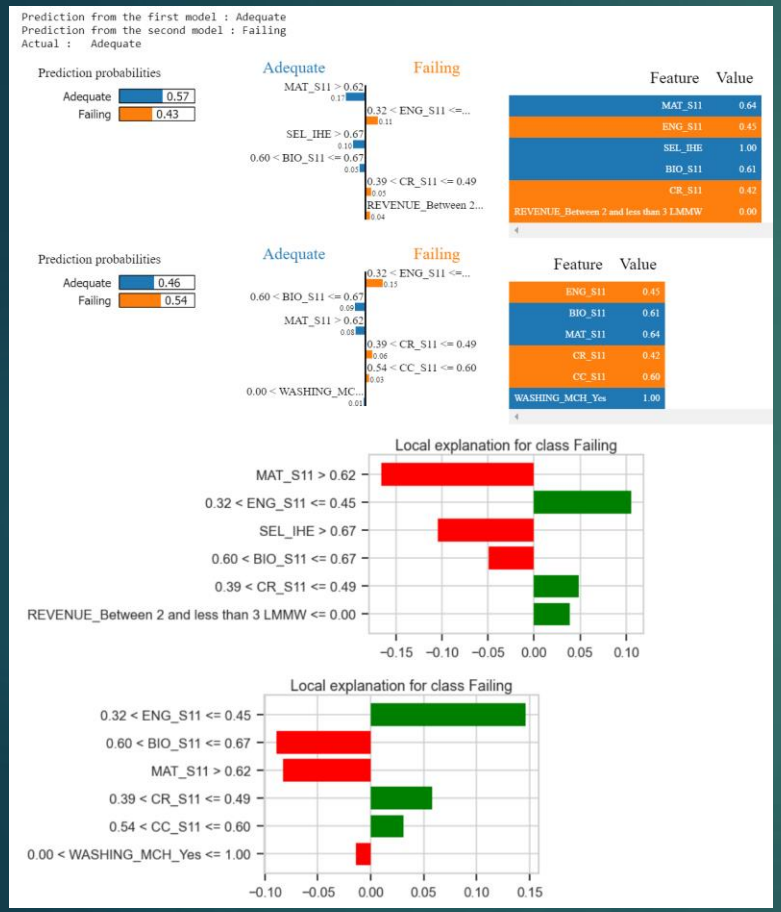
#4



- **Outstanding** HS test results
- In a program at a university with a **low-income** student population
- Both models predictions **in agreement** with the actual result
- Performance: **Adequate**

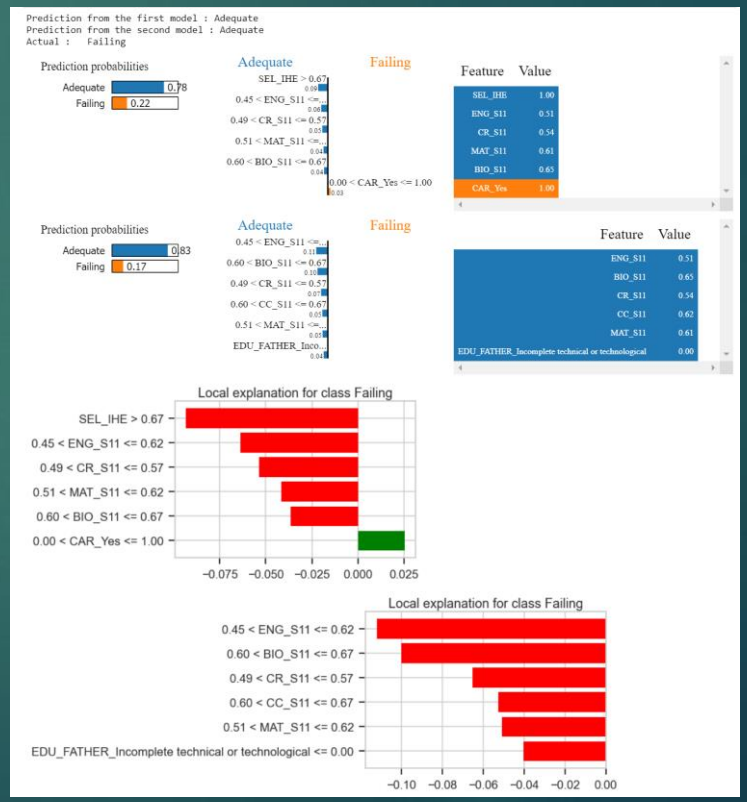
Use cases (continued)

#5



- **mixed** HS test results
- In a program at a university with a **low-income** student population
- The first model **failed** to predict the actual result
- The second model predictions **in agreement** with the actual result
- Performance: **Adequate**

#6

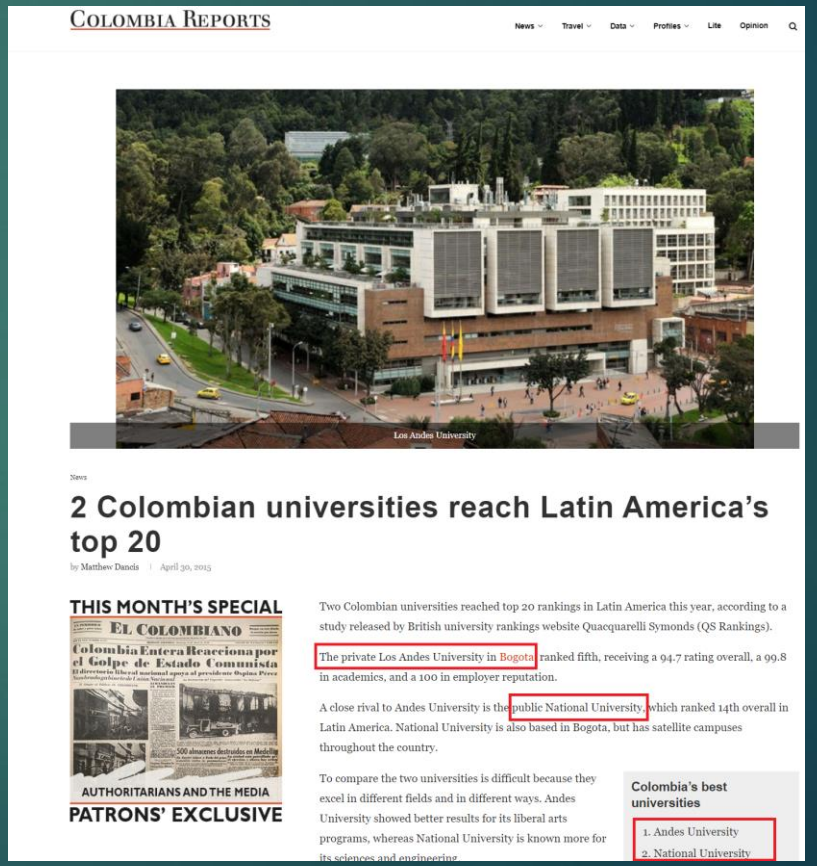


- **outstanding** HS test results
- In a program at a university with a **low-income** student population
- Both models did not predict the actual result
- Performance: **Failed**

For Future Engineering Students

Best		Worst
Model #1		
Civil Engineering		Civil Engineering
UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C.		UNIVERSIDAD COOPERATIVA DE COLOMBIA-BOGOTÁ D.C.
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		UNIVERSIDAD LA GRAN COLOMBIA-BOGOTÁ D.C.
UNIVERSIDAD DEL NORTE-BARRANQUILLA		CORPORACION UNIVERSIDAD DE LA COSTA, CUC-BARRANQUILLA
Industrial Engineering		Industrial Engineering
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		UNIVERSIDAD SIMON BOLIVAR-BARRANQUILLA
UNIVERSIDAD ICESI-CALI		UNIVERSIDAD DE PAMPLONA-PAMPLONA
UNIVERSIDAD DISTRITAL"FRANCISCO JOSE DE CALDAS"-BOGOTÁ D.C.		POLITECNICO GRANCOLOMBIANO-BOGOTÁ D.C.
Chemical Engineering		Chemical Engineering
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		FUNDACION UNIVERSIDAD DE AMERICA-BOGOTÁ D.C.
UNIVERSIDAD DE LA SABANA-CHIA		FUNDACION UNIVERSIDAD DE BOGOTA"JORGE TADEO LOZANO"-BOGOTÁ D.C.
UNIVERSIDAD DE CARTAGENA-CARTAGENA		
Mechanical Engineering		Mechanical Engineering
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		UNIVERSIDAD INDUSTRIAL DE SANTANDER-BUCARAMANGA
Model #2		
Civil Engineering		Civil Engineering
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		UNIVERSIDAD COOPERATIVA DE COLOMBIA-BOGOTÁ D.C.
UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C.		UNIVERSIDAD LA GRAN COLOMBIA-BOGOTÁ D.C.
UNIVERSIDAD DEL NORTE-BARRANQUILLA		CORPORACION UNIVERSIDAD DE LA COSTA, CUC-BARRANQUILLA
Industrial Engineering		Industrial Engineering
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		UNIVERSIDAD DE PAMPLONA-PAMPLONA
UNIVERSIDAD ICESI-CALI		UNIVERSIDAD SIMON BOLIVAR-BARRANQUILLA
UNIVERSIDAD DISTRITAL"FRANCISCO JOSE DE CALDAS"-BOGOTÁ D.C.		FUNDACION UNIVERSITARIA TECNOLOGICO COMFENALCO - CARTAGENA -CARTAGENA
		POLITECNICO GRANCOLOMBIANO-BOGOTÁ D.C.
Chemical Engineering		Chemical Engineering
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		FUNDACION UNIVERSIDAD DE AMERICA-BOGOTÁ D.C.
UNIVERSIDAD DE LA SABANA-CHIA		FUNDACION UNIVERSIDAD DE BOGOTA"JORGE TADEO LOZANO"-BOGOTÁ D.C.
UNIVERSIDAD DE CARTAGENA-CARTAGENA		
Mechanical Engineering		Mechanical Engineering
UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.		UNIVERSIDAD INDUSTRIAL DE SANTANDER-BUCARAMANGA

- Minimal disagreement between models
- Ranking in agreement with the official report



COLOMBIA REPORTS

News Travel Data Profiles Lite Opinion

Los Andes University

News

2 Colombian universities reach Latin America's top 20

by Matthew Dancis | April 30, 2015

THIS MONTH'S SPECIAL

EL COLOMBIANO

Colombia Entera Reacciona por el Golpe de Estado Comunista

AUTHORITARIANS AND THE MEDIA

PATRONS' EXCLUSIVE

Two Colombian universities reached top 20 rankings in Latin America this year, according to a study released by British university rankings website Quacquarelli Symonds (QS Rankings).

The private Los Andes University in Bogota ranked fifth, receiving a 94.7 rating overall, a 99.8 in academics, and a 100 in employer reputation.

A close rival to Andes University is the public National University, which ranked 14th overall in Latin America. National University is also based in Bogota, but has satellite campuses throughout the country.

To compare the two universities is difficult because they excel in different fields and in different ways. Andes University showed better results for its liberal arts programs, whereas National University is known more for its sciences and engineering.

Colombia's best universities

1. Andes University
2. National University

Recommendations

Recommendations to university officials:

- System to flag incoming students at risk
  - Factors affecting students' success: **HS test scores, SEL_IHE**
- Additional help to these students
  - **Tutoring**
  - **Smaller classes, etc.**
- SEL_IHE: a difficult one to tackle
- Invest in researching what factors drive SHE_IHE influence on students' success

Recommendations (continued)

Recommendations to future engineering students:

- High ranked Programs/Universities – a **good investment**
- Models **in agreement** with official ranking
- Bogota's **Los Andes University** and **National University**
  - The best programs in all fields
  - Consider moving from home, it is worth it

Conclusions:

Limitations of the model:

- Data might not be detailed enough
- The overall performance: not perfect
- Correlations of the features

Additional analysis suggested:

- Regression not Classification
- Add numerical features to describe Socio-Economic factors
- Consider global score for HS tests
- Update dataset with current data

Thank you!

Email: e.v.kazakova@gmail.com

GitHub: @sealaurel

LinkedIn: https://www.linkedin.com/in/elena-v-kazakova/

Appendix: data description

The dataset used in this project has been downloaded from the Mendeley Data Repository. The dataset has been published in ScienceDirect Elsevier publication **"Data in Brief," Volume 30, June 2020, 105537.**

Quote from the paper summary:

"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 Evaluation of Education (ICFES). The structure of the data allows us to observe the influence of social variables and the evolution of students' learning skills."

- The dataset has **44 dependent and independent variables**
- Each row represents a student
- The variables correspond to the student's

  personal information (primarily categorical)
  result obtained in the assessments (numerical)

The academic evaluation is recorded at two moments of the student's life

Between 2012 and 2014: the national standardized test scores at the final year of the high school **(Saber 11)**
In 2018: academic assessment in the final year of their professional training in Engineering **(SABER PRO)**

Appendix: data description (continued)

The dataset has **12411 unique records** with **44 columns**. There are no NULL values in any of the columns. The annotation to the fields and associated data

COD_S11: S11 test student identifier; 12411 unique values, no duplicates
Cod_SPro: SABER_PRO test student identifier; 12395 unique values, 16 duplicates

Categorical variables

GENDER: Student's gender; 2 unique values (F/M)
EDU_FATHER: Level of education of a student's father; 12 unique values
EDU_MOTHER: Level of education of a student's mother; 12 unique values
OCC_FATHER: Occupation of a student's father; 12 unique values
OCC_MOTHER: Occupation of a student's mother; 12 unique values
PEOPLE_HOUSE: Number of people in the household; 13 unique values
INTERNET: Does a household have an internet connection; 2 unique values (Yes/No)
TV: Does a household have a TV; 2 unique values (Yes/No)
COMPUTER: Does a household have a computer; 2 unique values (Yes/No)
WASHING_MCH: Does a household have a washing machine; 2 unique values (Yes/No)
MIC_OVEN: Does a household have a microwave oven; 2 unique values (Yes/No)
CAR: Does a household have a car; 2 unique values (Yes/No)
DVD: Does a household have a DVD player; 2 unique values (Yes/No)
FRESH: Does a household have access to freshwater; 2 unique values (Yes/No)
PHONE: Does a household have a landline phone; 2 unique values (Yes/No)
MOBILE: Does a student have a mobile phone; 2 unique values (Yes/No)
STRATUM: Colombian socio-economic class indicator; 7 unique values
SISBEN: Levels of Colombian welfare system; 6 unique values
REVENUE: Household income category; 8 unique values
JOB: An indicator if a student had a job; 4 unique values
SCHOOL_NAME: High School name, 3735 unique values
SCHOOL_NAT: High school Private or Public, 2 unique values
SCHOOL_TYPE: School Type, 4 unique values
UNIVERSITY: University Name, 134 unique values
ACADEMIC_PROGRAM: Engineering Program; 21 unique values

Appendix: data description (continued)

Numerical variables

- **SEL:** Student's Socio-Economic level; 4 unique values
- **SEL_IHE:** Average Socio-Economic level of the university/program a student attended; 4 unique level

High School academic assessment results

- **MAT_S11**: Mathematics
- **CR_S11**: Critical Reading
- **CC_S11**: Citizen Competencies
- **BIO_S11**: Biology
- **ENG_S11**: English

Engineering School academic assessment results

- **QR_PRO**: Quantitative Reasoning
- **CR_PRO**: Critical Reading
- **CC_PRO:** Citizen Competencies
- **ENG_PRO:** English
- **WC_PRO:** Written Communication
- **FEP_PRO:** Formulation of Engineering Projects
- **G_SC:** Global Score
- **PERCENTILE**: Percentile
- **Quartile**: Quartile 2ND_DECILE: Second Decile