**Tratamento da base**

$ **COD\_ALUNO:** os alunos 176525 e 265722 aparecem duas vezes na base, porém não parece haver inconsistências, pois parecem estar matriculados em duas disciplinas diferentes.

$ **COD\_UNIDADE:**

$ **COD\_TURMA:** 58 valores de unidades.

$ **SEXO**: Dois sexos (1 e 2)

$ **ESTADO\_CIVIL:** 9 valores para variável estado civil

**$ DATA\_NASCIMENTO:**

**$ CIDADE\_ALUNO:** Alguns campos aparecem números ao invés de nomes.

**$ UF\_ALUNO:** Aparecem 4 campos missing.

**$ CEP\_ALUNO:** Alguns campos aparecem com erros.

**$ COD\_CURSO:** Variável nominal

**$ NOME\_CURSO:** Variável nominal

**$ TURNO\_CURSO:** Variável nominal

**$ DURACAO\_CURSO:**

**$ DATA\_INGRESSO:** Variável nominal

**$ SITUCAO\_MATRICULA:** Variável nominal

**$ DATA\_CONCLUSAO:** Variável nominal

**$ PERIODOS\_TRANCADOS:**

**$ CH\_TOTAL\_MATRIZ:**

**$ NR\_TOTAL\_DISCIPLINAS:** apresenta dois outliers com valores 0, e 1 como total de disciplinas

**$ CH\_APROVADA:** Alguns campos apresentam valores menores que o CH\_TOTAL\_MATRIZ (1.449 observações).

**$ CH\_A\_CURSAR:** A variável não aparece na base.

**$ NR\_DISC\_APROVADA:** Alguns campos apresentam valores menores que o NR\_TOTAL\_DISCIPLINAS (1.449 observações).

$ **T\_IDADE:** apresenta duas observações com valores 2, e 3. Provavelmente são erros de preenchimento de cadastro.

$ **NOTA:** A variável apresenta mínimo de 6,19 e máximo de 9,83. Apresenta baixa variância, talvez por causa da base apresentar apenas os alunos que se formaram.

$ **NOTA\_ENEM:**

$ **POSSUI\_ENEM:**

$ **NRO\_REPRO\_ACO:** apresentam 6.487 observações com missing.

$ **NRO\_REPRO\_NORMAL:** apresentam 6.487 observações com missing.

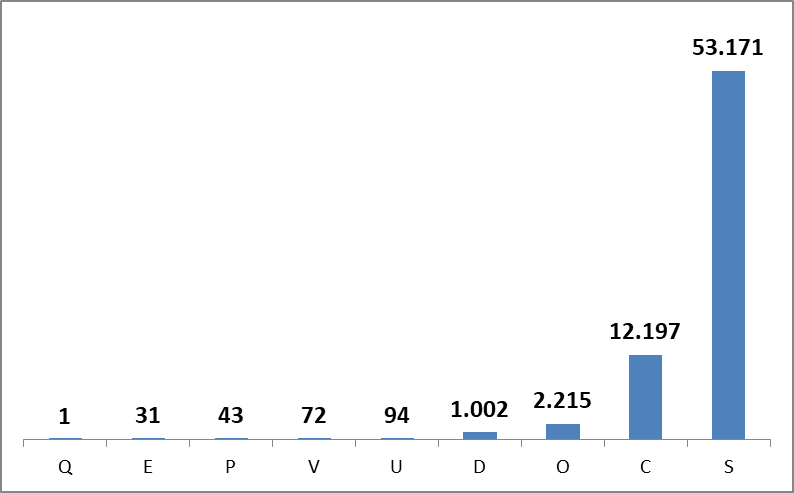
$ **NRO\_TOTAL\_REPRO:** apresentam 6.487 observações com missing.

$ **SEMESTRES\_CURSADOS:** aparecem alguns outliers com número de semestres cursados menores do que a quantidade de semestres de duração, porém como o status dos alunos são todos formados, é evidência de que sejam outliers.

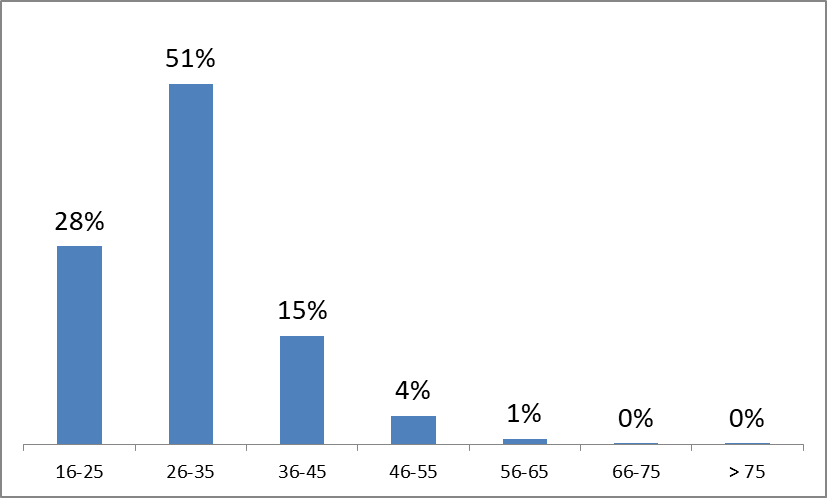
$ **POSSUI\_FIES:** Variável nominal

**Análise Demográfica Prévia**

Abaixo a distribuição de frequência de alunos por estado civil. A grande maioria dos alunos são solteiros, e a segunda maior parte são casados.

  
Gráfico de frequência da variável ESTADO\_CIVIL

Abaixo a distribuição de frequência por faixa etária. Quase 80% dos alunos são adultos até 35 anos. Metade dos alunos se encontram na faixa de 25 à 35 anos, 28% tem até 25 anos. Somente 5% dos alunos possuem mais que 45 anos.

  
Distribuição de frequência por faixa de idade em percentual sobre o total (excluindo dois outliers)

Abaixo é possível ver uma tabela com a distribuição de alunos por estado civil. A tabela mostra o percentual de alunos que possuem FIES e que não possuem, sexo 1 e sexo 2 por estado civil. Dos alunos, a maioria é solteira, não possui FIES, do sexo 2. Em geral 62% dos alunos não possui FIES, e 61% são do sexo 2.

  
Distribuição de frequência por estado civil.

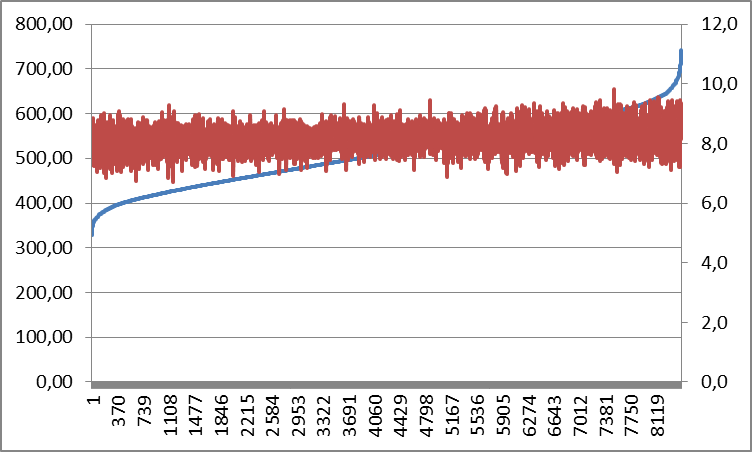
A segmentação estado civil não apresenta grande variação entre as variáveis de performance, como pode ser visto na tabela abaixo. A nota média, a nota média do ENEM não apresentam grande variação entre os tipos de estado civil. O mesmo acontece para total de reprovações, reprovações aco, e reprovações normais.

Tabela por estado civil demonstrando baixa variância em variáveis de performance estudantil.

Na tabela abaixo é possível ver a análise segmentada para faixa de idade dos alunos. Vale notar que a nota média não apresenta grande variação entre as faixas, mas há uma evidente correlação entre idade e nota, alunos em faixas etária mais elevada possuem performance levemente inferior a alunos em faixa etária inferiores.

  
Tabela de faixa etária com quantidade de alunos e nota média.

As notas de ENEM e nota da faculdade apresentam baixa correlação de 34%. No gráfico abaixo a plotagem das curvas de nota da faculdade em vermelho, e da nota do ENEM em azul. O eixo da esquerda representa as notas do ENEM, variando de 0 a 800, e no eixo da direita representa a nota média da faculdade, variando de 0 a 10.

  
Gráfico para correlação entre nota curso faculdade (vermelho) vs nota do ENEM (azul)

A variável geográfica UF do aluno mostra que alunos do DF, AM, RR, RN, PB não fizeram ENEM. A maior nota média do ENEM é dos alunos de SE, enquanto que as menores notas médias do ENEM são dos alunos do PI. A tabela demonstra o “share de alunos” por UF. A tabela abaixo mostra que a correlação entre nota do ENEM e performance universitária não são tão correlacionadas, o total de reprovações não apresenta grande correlação com as notas médias do ENEM, na verdade apresentam correlação negativa no valor de -28%. O mesmo acontece entre as notas médias da faculdade, e as notas médias do ENEM, onde a correlação é de apenas 46%. A variação dos semestres cursados é baixa, e em geral sempre maior do que a duração média dos cursos. Os dados evidenciam que em geral os alunos demoram mais do que o esperado para se formar, e isso independe da UF em que residem.

  
tabela de segmentação por UF do aluno

**Metodologia de cálculo**

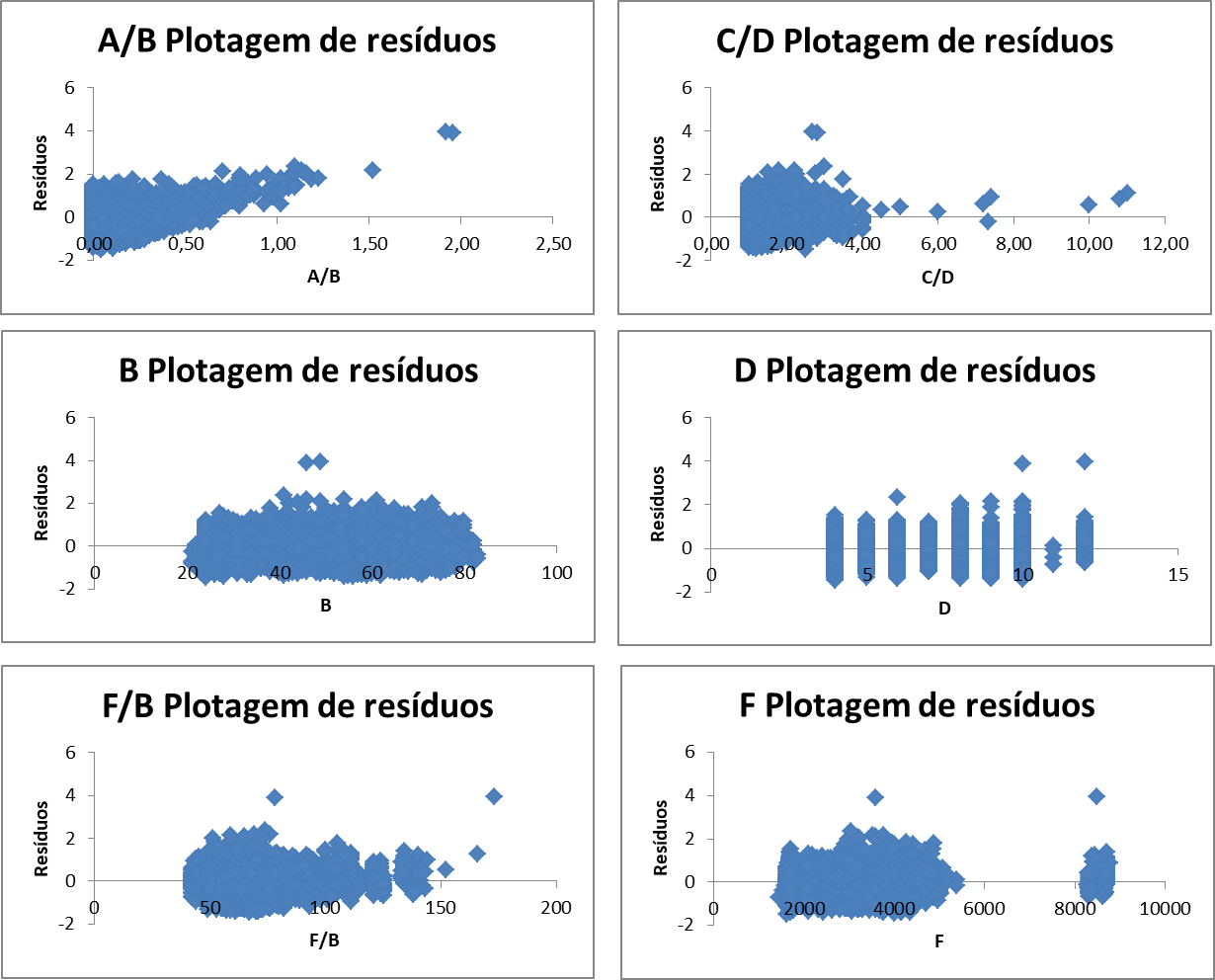
O modelo utilizado para cálculo de probabilidade de previsão de formatura é uma simples regressão linear, no caso a variável resposta não é uma probabilidade em si, mas um score. Como foi descrito nos pontos de melhoria, seria interessante possuir uma variável resposta binomial, onde na base teria não somente os alunos que se formaram, mas também aqueles que não conseguiram se formar. Deve-se observar que é possível criar modelos muito mais descritivos, com mais variáveis, e um maior poder descritivo, mas dado o tempo limitado para a criação do modelo, foi selecionado um modelo simples para se obter um resultado quantitativo e qualitativo dado o curto período de tempo.  
 Premissas utilizadas para o modelo de regressão:

* A variável SEXO não foi utilizada no modelo, pois estatisticamente não há diferença em níveis de inteligência entre homens e mulheres.
* A variável idade não foi utilizada no modelo, pois não apresenta variância em NOTA média do aluno.
* A variável ESTADO\_CIVIL não foi utilizada no modelo, pois não apresentou variância suficiente em variáveis de desempenho: NOTA, NOTA\_ENEM, NRO\_REPRO\_ACO, NRO\_TOTAL\_REPRO, NRO\_REPRO\_NORMAL, SEMESTRES\_CURSADOS.
* A variável NOTA foi utilizada como proxy de probabilidade de conclusão do curso. Foi utilizada como variável dependente na regressão linear.
* Foram retiradas as observações que continham outliers como foi descrito na seção **Tratamento de Base**, foram mantidas somente as observações (alunos) sem missing, e sem valores incorretos.
* As variáveis NR\_DISC\_APROVADA e CH\_APROVADA apresentam a mesma quantidade coincidente de observações menores que as variáveis NR\_TOTAL\_DISCIPLINA e CH\_TOTAL\_MATRIZ. Essas observações foram tratadas como outlier, pois o metadados de variáveis não apresenta descrição adequada, ou critério do motivo dos valores aprovados de disciplina e carga horária serem menores do que o total requerido para formação. Como se trata de apenas 1.449 observações, não afetará a base para o modelo.
* As variáveis auxiliares originais da base de dados utilizadas no modelo foram:
  + NOTA: variável dependente (resposta)
  + NRO\_REPRO\_NORMAL (A)
  + NR\_TOTAL\_DISCIPLINAS (B)
  + SEMESTRES\_CURSADOS (C)
  + DURACAO\_CURSO (D)
  + PERIODOS\_TRANCADOS (E)
  + CH\_TOTAL\_MATRIZ (F)
* As variáveis de performance criadas a partir das variáveis originais da base de dados, a serem utilizadas no modelo como variáveis independentes foram:
  + **A/B:** Esta variável é o número de reprovações normais pelo número total de disciplinas. É um indicador de desempenho. O racional para usar esta variável é detectar um percentual de reprovações pela quantidade de disciplinas que seja crítico para que o aluno se forme.
  + **B:** Número total de disciplinas foi utilizado, pois é um indicador natural de dificuldade. Quanto maior a quantidade de disciplinas de um curso, maior a dificuldade em se formar.
  + **C/D:** Quantidade de semestres cursados pela duração total do curso. É um indicador de desempenho que demonstra um valor crítico entre o tempo que o aluno está no curso, em comparação ao total de semestres em que o curso deveria ser terminado.
  + **F:** É um indicador natural de dificuldade em se terminar um curso. Quanto maior a carga horária de um curso, maior a dificuldade.
  + **F/B:** Indicador de dificuldade criado seguindo o racional de desempenho entre as variáveis F, e B. A carga horária por disciplina é correlacionada com a dificuldade em cursar as disciplinas. Quanto maior a carga por disciplina, maior a dificuldade.
* A variável PERIODOS\_TRANCADOS foi rejeitada pelo p-valor em regressão prévia, e por isso não foi colocada no modelo.

Na tabela abaixo segue os resultados da regressão linear, utilizando o Excel.

  
Resultados obtidos pela regressão linear (via análise de dados do Excel – Regressão)

Os betas do modelo se encontram na coluna Coeficientes, e os nomes das variáveis se encontram na coluna à esquerda. Os valores de p-valor são todos abaixo de 0,5%. O modelo apresentou correlação de 45%, erro relativo médio de -0,01%, erro absoluto médio de 33,25%. Nos gráficos abaixo é possível ver a distribuição dos resíduos para cada uma das variáveis. Apenas a variável A/B apresenta leve tendência entre os resíduos, entretanto testes de normalidade demonstram que todas as variáveis são aceitas pelo modelo de regressão linear. A distribuição dos resíduos das outras variáveis são claramente normais.



**Observações:** É possível realizar modelos melhores com mais variáveis preditivas, entretanto devido ao curto tempo para realizar o tratamento da base, entender as variáveis, e propor um modelo razoável, este modelo em questão apresenta boa performance preditiva, apesar de sua simplicidade. Um modelo de regressão logística seria mais adequado para prever sucesso em se formar do que a nota média da faculdade. O fato de constar na base apenas os alunos que se formaram prejudica a disponibilidade de modelos a serem aplicados, seria interessante utilizar como variável resposta uma proxy de sucesso/fracasso (distribuição binomial). As notas médias apresentam baixa dispersão de apenas 4,68%.

O modelo foi desenvolvido no EXCEL, pois apresentava baixo numero de observações para o tratamento de base, e desenvolvimento. O fator prazo também foi preponderante, já que softwares em formato de planilha (excel, minitab, sas guide, etc.) apresentam certa facilidade em tratar linhas e colunas quando o número de variáveis, e observações é baixo.

**Pontos de Melhoria**

A base disponibilizada apresenta a variável SITUACAO\_MATRICULA, na base constam apenas os alunos formados. Uma metodologia bastante comum seria criar um score para os alunos através de uma regressão logística, para tal a base deveria conter também os alunos que largaram o curso e não conseguiram se formar, assim se teria uma variável binomial, com 1 para sucesso (formado), e 0 para fracasso (não formou). Com a regressão logística é possível utilizar variáveis nominais como, por exemplo, TURNO\_CURSO, DURACAO\_CURSO, NOME\_CURSO, POSSUI\_FIES, entre outras.   
 O metadados de variáveis poderia ser mais descritivo, o data scientist teria que adivinhar pelo bom senso algumas coisas, como por exemplo, a variável DURACAO\_CURSO, deveria estar descrito que a duração é em semestres.   
 Poderia haver mais variáveis descritivas de segmentação na base. Por exemplo, poderia haver uma variável auxiliar de grande área, separando os cursos em Exatas, Humanas, Biológicas. Outra variável auxiliar se os cursos são tecnólogos, bacharelado, licenciatura.

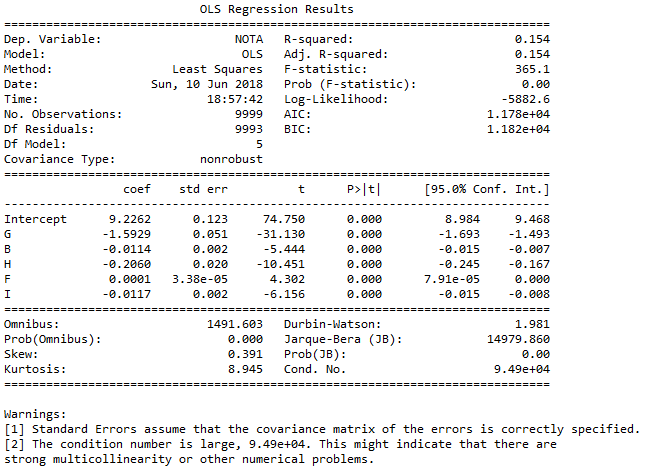
Além do modelo preditivo eu utilizaria variáveis econômicas e financeiras, por exemplo, valor de mensalidade paga, valor de taxa por disciplina reprovada, valor por semestre extra, renda familiar, IPCA, ciclo econômico, etc. O racional seria utilizar a correlação entre performance estudantil e situação econômica e financeira do país. Quando a crise se intensifica o cidadão privilegia seus custos básicos (moradia, alimentação, vestuário).

Procuraria variáveis de segmentação além das mais corriqueiras (eg. sexo, idade, UF), procuraria variáveis como renda familiar média, salário, se a pessoa está empregada ou não, se fez o ensino médio em escolas particulares ou públicas, se possui ensino técnico, etc.

Além do modelo de regressão linear eu gostaria de ter feito uma segmentação via K-Means ou KNN para realizar estudo das variáveis, quais conseguem caracterizar melhor os modelos.

**ANEXO: Modelo Implementado em Python**

Resultados do modelo



**Abaixo o código em Python**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import statsmodels.formula.api as sm

#lê o arquivo csv a partir do caminho

df = pd.read\_csv(r'C:\Users\OsvaldoLuiz\Desktop\formatura.csv',

sep=';', encoding='latin-1',

delimiter=None, header='infer', names=None, index\_col=None,

usecols=None, squeeze=False, prefix=None, mangle\_dupe\_cols=True,

dtype=None, engine=None, converters=None, true\_values=None,

false\_values=None, skipinitialspace=False, skiprows=None,

nrows=None, na\_values=None, keep\_default\_na=True, na\_filter=True,

verbose=False, skip\_blank\_lines=True, parse\_dates=False,

infer\_datetime\_format=False, keep\_date\_col=False, date\_parser=None,

dayfirst=False, iterator=False, chunksize=None, compression='infer',

thousands=None, decimal=b'.', lineterminator=None, quotechar='"',

quoting=0, escapechar=None, comment=None, dialect=None,

tupleize\_cols=None, error\_bad\_lines=False,

warn\_bad\_lines=True, skipfooter=0, doublequote=True,

delim\_whitespace=False, low\_memory=True, memory\_map=False,

float\_precision=None)

#imprime o total de linhas do arquivo

print("Total rows: {0}".format(len(df)))

#lista os nomes das colunas do arquivo

#print(list(df))

#imprime o formato do arquivo

#print(df.shape)

#imprime as 5 primeiras linhas do cabeçalho

#print(df.head())

#imprime o arquivo

#print(df)

#imprime a descrição das colunas do arquivo

#print(df.describe())

#print(df.info())

#imprime os valores únicos de todas as colunas

#for col in df:

# print(df[col].unique())

df=df.dropna(subset=['NRO\_REPRO\_NORMAL',

'NR\_TOTAL\_DISCIPLINAS','SEMESTRES\_CURSADOS',

'DURACAO\_CURSO','DURACAO\_CURSO',

'PERIODOS\_TRANCADOS','CH\_TOTAL\_MATRIZ'])

df['A'] = df['NRO\_REPRO\_NORMAL']

df['B'] = df['NR\_TOTAL\_DISCIPLINAS']

df['C'] = df['SEMESTRES\_CURSADOS']

df['D'] = df['DURACAO\_CURSO']

df['E'] = df['PERIODOS\_TRANCADOS']

df['F'] = df['CH\_TOTAL\_MATRIZ']

df['G'] = df['A']/df['B']

df['H'] = df['C']/df['D']

df['I'] = df['F']/df['B']

amostra = df.sample(10000,replace='False')

#print(amostra)

lm = sm.ols(formula=

'NOTA ~ G + B + H + F + I', data=amostra).fit()

print(lm.summary())