Estudo do Perfil Discente de Graduação na UFABC

Gabriel P. de Carvalho, Gabriel S. Mancini, Rafael Z. Hernandez, Renato M. da Silva

[1]Fundação Universidade Federal do ABC (UFABC)
Santo André – SP – Brazil

pitalig@gmail.com, mancinigabriel@hotmail.com

rafael_zayat@hotmail.com, renato_morassi@hotmail.com

Abstract. *This project has as objective the analysis of the data collected through a student profile survey enrolled in 2018 at UFABC, aiming to find a relationship between the demographic and academic-related features. This analysis was done through statistical and machine learning methods. After preprocessing the data, we searched for correlations and selected some variables that presented some correlation with the academic performance or had an interesting demographic feature and assembled them on clusters. Analyzing the defined clusters we did not identify data that proved our hypothesis that it was possible to correlate demographic data directly to the academic performance.*

Resumo. *Este projeto tem como objetivo analisar os dados coletados através da pesquisa de perfil discente dos alunos matriculados em 2018 na UFABC, buscando relações entre as variáveis demográficas da pesquisa e o rendimento acadêmico dos alunos dentro da amostra. Essa análise foi feita através de métodos estatísticos e de aprendizado de máquina, após o pré processamento dos dados foram procuradas correlações entre as variáveis, selecionadas variáves que apresentavam correlação, mesmo que mínima, com o rendimento ou que apresentavam caráter demográfico interessante e com elas montamos "clusters"de amostras. Analisando os clusters definidos não identificamos dados que provassem nossa hipótese de que era possível correlacionar dados demográficos diretamente com o rendimento acadêmico dos alunos.*

1. Introdução

Este trabalho nasceu da hipótese de que as informações demográficas dos alunos deveriam afetar diretamente o rendimento acadêmico (CR) dos alunos da UFABC, hipótese essa que nos parecia direta ao início do trabalho, para tentar provar isso usamos os dados fornecidos pela "Pesquisa de Perfil Discente de Graduação"da UFABC. A proposta inicial era identificar as maiores correlações demográficas com o rendimento e tentar prever através delas o rendimento esperado do aluno ou o classificar em um grupo que se relacionasse com um rendimento acadêmico.

2. Dataset e Tratamento Prévio dos Dados

2.1. Dataset - "Pesquisa de Perfil Discente da Graduação"

O trabalho foi todo desenvolvido com base no dataset da "Pesquisa de Perfil Discente da Graduação"[UFABC 2018], pesquisa realizada anualmente pela UFABC no momento

da matrícula de um quadrimestre específico. A pesquisa é composta por perguntas que abordam aspectos acadêmicos, pessoais e socioeconômicos, além de uma avaliação da infraestrutura da Universidade, com o intuito de acompanhar a evolução do perfil dos discentes ao longo dos anos e coletar informações para subsidiar a avaliação de políticas institucionais. O dataset é composto de 323 features, que envolvem diversas perguntas com respostas definidas, como por exemplo ”Qual a sua principal fonte de sustento?”e perguntas com respostas abertas. Na pesquisa de 2018 foram colhidas 9929 respostas.

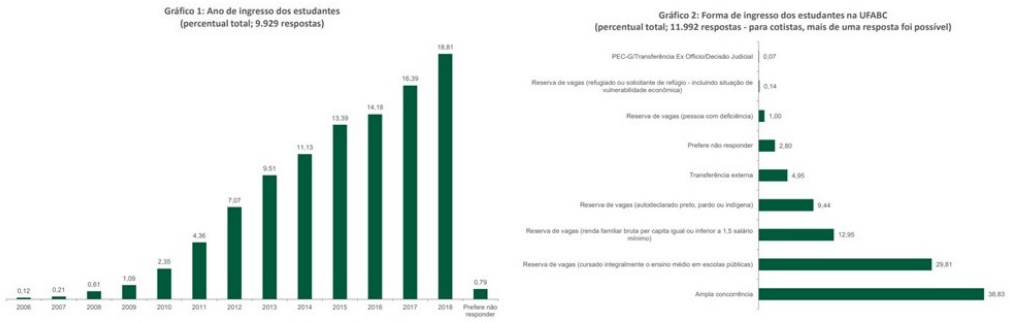

Figura 1. Exemplo de informações colhidas através da pesquisa

2.2. Tratamento dos Dados

Antes de utilizar os dados obtidos e aplicá-los no algoritmo, é necessário observar analíticamente o dataset e filtrar manualmente dados que provavelmente não contribuirão para o desenvolvimento do modelo ou que fornecerão resultados incorretos dado sua natureza e o método de análise utilizado. Inicialmente, analisamos as features do dataframe e escolhemos as que poderiam ser transformadas em valores númericos e ter sua correlação estudada, tendo essas features definidas foi feito o tratamento delas para um formato numérico e foram descartadas as amostras que continham respostas em branco nessas features escolhidas. Após o tratamento dos dados restaram 6346 amostras consideradas úteis para o nosso estudo.

2.3. Features

Após o tratamento, escolhemos nove features:

- Idade
- Coeficiente de Rendimento (CR)
- Horas de Permanência na Universidade
- Tempo de Deslocamento até a Universidade
- Quantidade de Trancamentos Totais de Matrícula Realizados
- Quantidade de Disciplinas Reprovadas
- Ano de Ingresso
- Renda Familiar Total
- Quantidade de Pessoas na Família

Para que os dados estivessem minimamente normalizados, transformamos algumas dessas featuras, ao invés de usarmos o ano de ingresso, transformamos na quantidade de anos em que o aluno estava matriculado na UFABC e calculamos a renda per capita em função de 1000R$, dividindo a renda familiar total pelo tamanho da família.

No final ficamos com 8 features para serem analisadas:

- Idade
- Coeficiente de Rendimento (CR)
- Horas de Permanência na Universidade
- Tempo de Deslocamento até a Universidade
- Quantidade de Trancamentos Totais de Matrícula Realizados
- Quantidade de Disciplinas Reprovadas
- Quantidade de Anos do Aluno na UFABC
- Renda Familiar Per Capita/Mil

3. Metodologia

3.1. Análise de Correlações

Tendo a base com as 8 features tratadas, analisamos como cada feature se correlacionava com o rendimento usando coeficientes de correlação de Pearson, representamos essas correlações através de uma matriz onde a intensidade da cor representa correlação alta ou baixa e os tons entre azul e vermelho representam correlação direta ou inversa. Foi identificado que as variáveis que representam a quantidade de reprovações, idade e anos de UFABC apresentavam alguma correlação com o rendimento, mesmo que baixa, e que o tempo de deslocamento até a UFABC e a renda per capita apresentavam correlação baixíssima ou quase 0 com o coeficiente de rendimento. A matriz é apresentada a seguir.

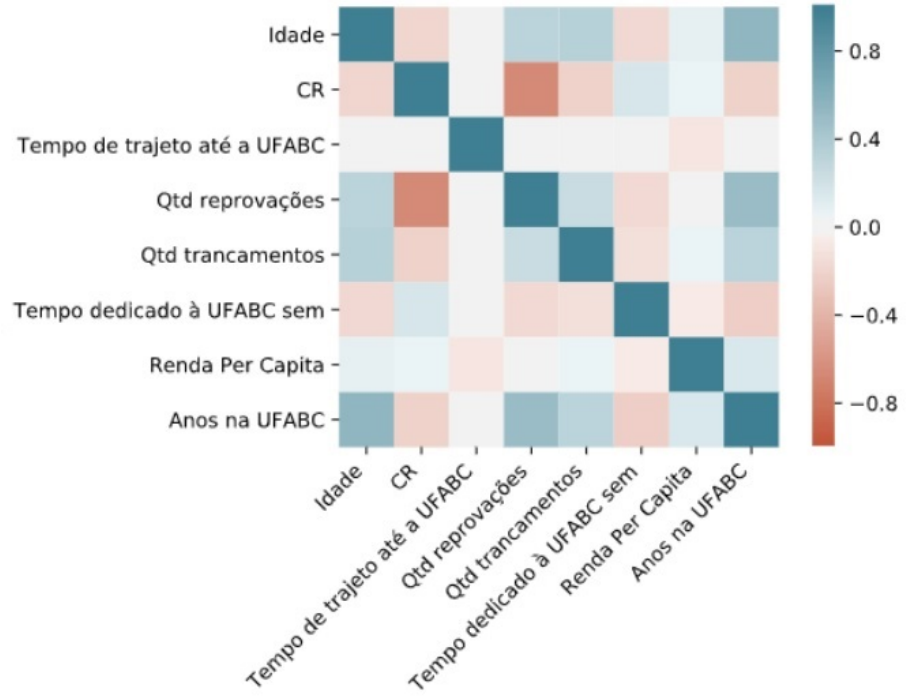

Figura 2. Matriz de correlação de Pearson

Após observar as correlações foi possível selecionar algumas features para considerarmos na clusterização de acordo com as correlações, pela correlação mantivemos a idade, quantidade de reprovações e anos desde a matrícula da UFABC além dessas variáveis optamos por manter na análise a renda per capita, pois o estudo visa utilizar

variáveis demográficas e entendemos que essa variável poderia ser interessante na análise. As observações são dadas em coeficientes de acordo com a matriz. Abaixo duas tabelas, uma usada para mostrar como os coeficientes são representados numericamente e outra mostrando as correlações das variáveis escolhidas para a análise.

Tabela 1. Correlações encontradas

Feature A	Feature B	Correlação
CR	CA	0.960909
Idade	Anos desde a matrícula na UFABC	0.533560
Reprovações	Anos desde a matrícula na UFABC	0.506287
CR	Reprovações	-0.663620

Tabela 2. Variáveis utilizadas

Feature A	Feature B	Correlação
CR	Reprovações	-0.663620
CR	Anos desde a matrícula na UFABC	-0.214284
CR	Idade	-0.197127
CR	Renda per capita	0.055675

3.2. Normalização e Redução de Dimensionalidade

Para aproximar as escalas dos dados utilizados, foi usada uma função de normalização da biblioteca scikitlearn, normalizando pelo valor médio dos vetores, além disso utilizamos a técnica de PCA (Principal Component Analysis) para reduzir a dimensão das 4 features para 2 (PCA = 2).

3.3. Método Cotovelo

Com o grupo de features definido, as variáveis normalizadas e a dimensão ajustada, o próximo passo consiste na aplicação do método KMeans para clusterização, mas para isso é necessário definir o número ótimo de clusters para o modelo, para definirmos isso utilizamos o ”Método do Cotovelo”[Bholowalia 2014]. O método consiste na execução da clusterização utilizando KMeans com diversos números de clusters e então é analisado a variação da inércia do modelo em função da quantidade de clusters, o quantidade ideal de clusters está no ponto do gráfico que estiver mais distante da reta entre o primeiro e o

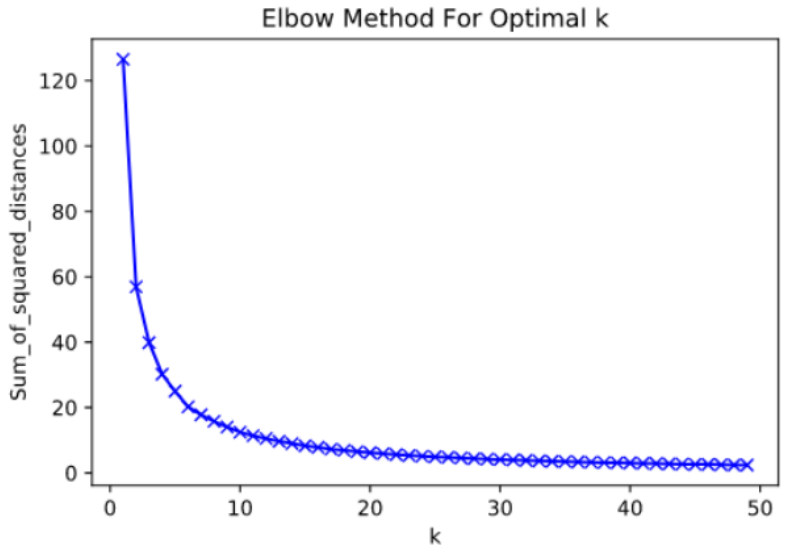

Figura 3. Gráfico da aplicação do ”Método Cotovelo”

último ponto do gráfico. Esse ponto marca o momento em que a diminuição da inércia do modelo começa a diminuir muito conforme a quantidade de clusters aumenta, fazendo com que não valha a pena fazer com que o modelo tenha mais clusters além daquilo. No nosso caso, encontramos o ponto ótimo em 8 clusters.

3.4. Clusterização utilizando KMeans

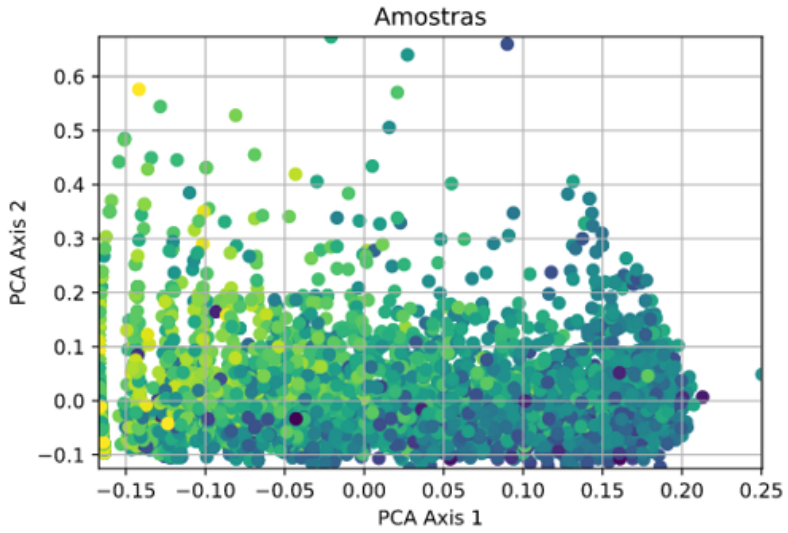

Figura 4. Distribuição das amostras após o PCA

Após definido o número ótimo de clusters, foi utilizado o método KMeans para agrupar os resultados em oito diferentes clusters, usando como base a variação na inércia do modelo. O algoritmo Kmeans agrupa os pontos de acordo com a distância entre eles, plotando pontos de acordo com a quantidade de clusters definidos e ajustando a distância deles às amostras até que as distâncias cobertas pelos grupos sejam o mais próximas possível. Podemos observar isso na distribuição dos dados e em como os dados foram agrupados nessa distribuição.

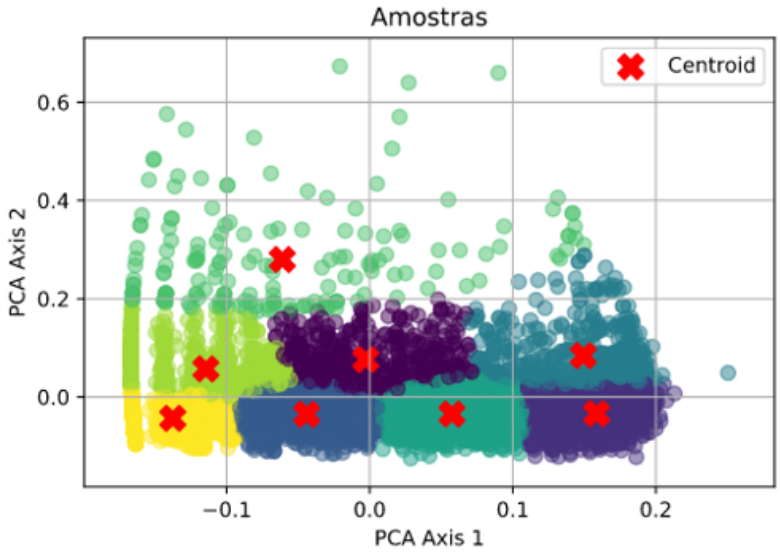

Figura 5. Distribuição dos 8 clusters e centroides nas amostras após o PCA

4. Análise dos Clusters

Após identificarmos os clusters e aplicarmos o algoritmo na base orginal, fizemos uma análise de como as médias das variáveis escolhidas se comportavam dentro dos cluester, procurando as suposições levantadas no início do trabalho e não conseguimos encontrar nenhuma relação direta através da análise dos valores dentro dos clusters.

Cluster	CR	Reprovações	Renda per Capita	Anos desde a matrícula
0	2.9587	0.2257	3062.9430	2.0700
1	1.7821	6.7736	1301.8426	4.5058
2	2.5182	2.1897	3563.7315	4.4634
3	2.7514	1.4633	9074.8588	2.8023
4	2.4055	1.8688	1127.8705	2.5987
5	2.8132	0.2756	1101.1896	0.7524
6	2.0771	4.5763	1372.4261	3.4375
7	1.9400	6.4364	4251.9962	5.9636

Figura 6. Tabela com as médias dos valores encontrados nos clusters.

Referências

Bholowalia, P., K. A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.735.7337&rep=rep1&type=pdf`. Acessado em: 2019-11-25.

UFABC (2018). Perfil discente de graduação. `http://propladi.ufabc.edu.br/informacoes/perfil`. Acessado em: 2019-11-25.