UNIVERSIDADE FEDERAL DE MINAS GERAIS

DCC212 – INTRODUÇÃO À CIÊNCIA DOS DADOS

Relacionando produtividade e investimento público em bolsas: uma análise dos cursos de pós-graduação da Universidade Federal de Minas Gerais

Relatório do Projeto Final

Ewerton S. Santos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG - Brasil
ewerton_dc@hotmail.com

Giovanni F. Martinelli
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG - Brasil
giofmartinelli@gmail.com

Gustavo R. A. Rodrigues
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG - Brasil
gustavorodrigues@dcc.ufmg.br

Rafael A. B. Perez
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG - Brasil
rafael.perez@dcc.ufmg.br

Vídeo: Acesse o vídeo do trabalho

1. Introdução e Motivação

A motivação para o nosso trabalho decorreu-se devido aos recentes questionamentos e bloqueios de verba do governo atual em relação às instituições federais. Tendo isso em mente, tivemos a ideia de explorar os relacionamentos entre investimento público e a produção acadêmica dos cursos de pós-graduação da Universidade Federal de Minas Gerais (UFMG).

Este trabalho tem por objetivo realizar uma avaliação do impacto dos investimentos financeiros sobre o volume da produção científica dos cursos de Pós-graduação da UFMG. Para isso, pretende-se observar como as variações nos valores investidos afetam os resultados de pesquisa ao longo dos anos.

A UFMG atualmente possui 75 cursos de pós-graduação acadêmicos, segundo avaliação quadrienal da CAPES. Para este trabalho foram selecionados os cursos de pós-graduação com modalidade de doutorado e mestrado entre os anos de 2008 e 2017.
Neste trabalho queremos entender a produtividade e tentar descobrir quais relações decorrem dos investimentos nos cursos.

2. Problema

Perguntas de Pesquisa

- Qual a relação entre o número de bolsas de mestrado e doutorado e o volume de produção por curso?
- Qual a influência da nota CAPES sobre o número de bolsas por curso?
- Qual a influência da nota CAPES sobre a produtividade de um programa?

3. Metodologia

Descrição da Base
Para realizar o trabalho foram utilizadas fontes de dados de domínio público, e com elas, geramos a nossa base. O grupo quis vivenciar a experiência e o desafio de criar a própria base. A seguir, descrevemos como são essas bases originais.

3.1. Dados Abertos da CAPES

3.1.1 Dados Abertos da CAPES: Produção Intelectual da Pós-Graduação

Descrição: Conjunto de dados referentes às produções artísticas, bibliográficas e técnicas dos programas de pós-graduação do Brasil no período compreendido entre 2004 e 2017. Este conjunto apresenta informações sobre:

1. Nome Do Programa De Pós-Graduação
2. Instituição De Ensino
3. Modalidade Do Programa De Pós-Graduação
4. Área Do Conhecimento
5. Ano
6. Titulo Da Produção
7. Nome Do Projeto
8. Modalidade Da Produção
9. Nome Do Autor

Fonte: Produção Intelectual da Pós-Graduação.

3.1.2 Dados Abertos da CAPES: Catálogo de Teses e Dissertações

Descrição: Conjunto de dados referentes a teses e dissertações dos programas de pós-graduação do Brasil a partir do final de 1987 até 2018. Este conjunto apresenta informações sobre:

1. Nome Do Programa De Pós-Graduação
2. Instituição De Ensino
3. Área Do Conhecimento
4. Ano
5. Nome Do Projeto
6. Título
7. Resumo
8. Modalidade (Tese Ou Dissertação)
9. Nome Do Autor
10. Nome Do Orientador

Fonte: Catálogo de Teses e Dissertações.

3.2. GeoCapes

Descrição: Conjunto de dados referentes às produções artísticas, bibliográficas e técnicas dos programas de pós-graduação do Brasil a partir do final de 1995 até 2018. Este conjunto apresenta informações sobre:

1. Nome Do Programa De Pós-Graduação
2. Instituição De Ensino
3. Área Do Conhecimento
4. Programa De Fomento
5. Bolsas De Iniciação Científica
6. Bolsas De Mestrado
7. Bolsas De Doutorado

Fonte: GeoCapes

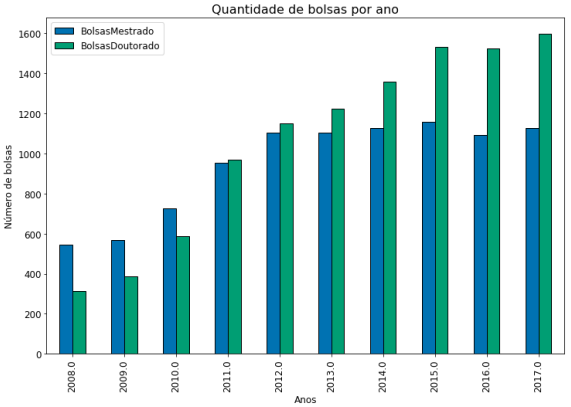
Limpeza dos dados
Foi feita uma triagem inicial com processamento, unificação e transposição de dados referentes aos programas de Pós-graduação da UFMG, tendo como fonte os dados abertos fornecidos pela plataforma Sucupira. Como resultado obtivemos uma planilha que agregou as informações relevantes à nossa pesquisa.

4. Resultados e análises

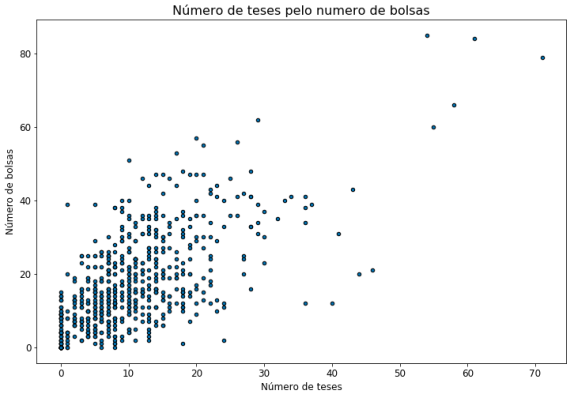
Caracterização – Análise Exploratória

Gráficos: Bolsas

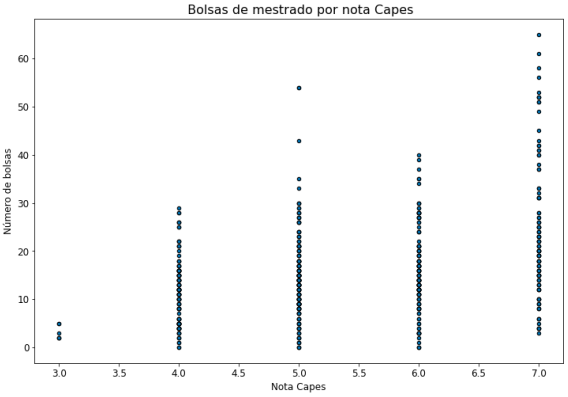
Bolsas por ano



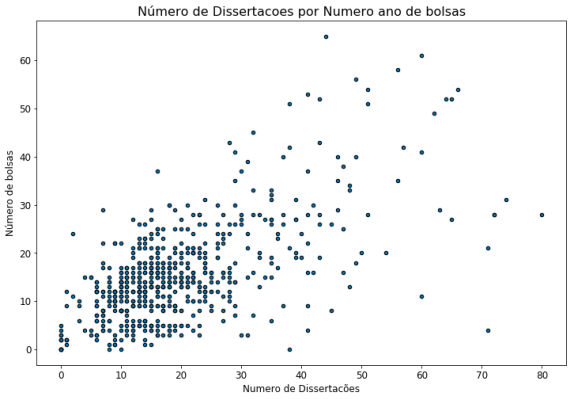
Bolsas pelo número de teses



Bolsas de mestrado pelo nível CAPES

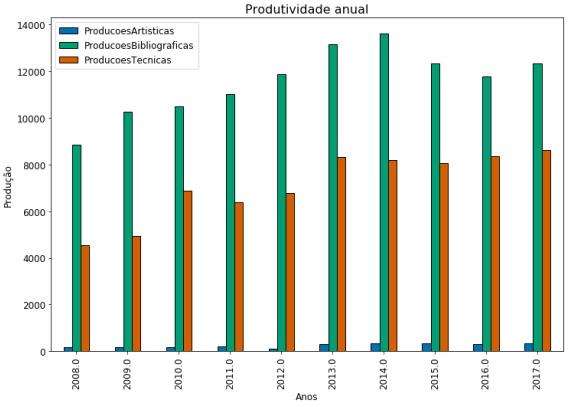


Bolsas pelo número de dissertações

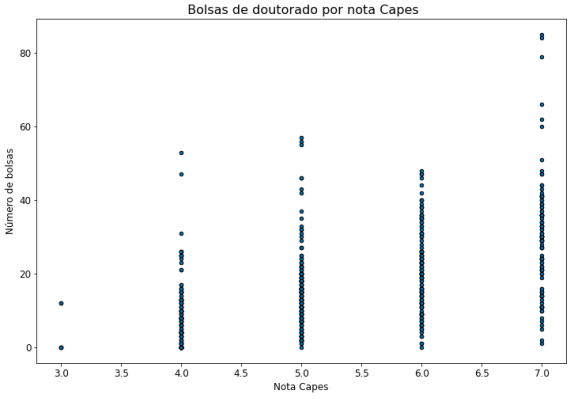


Gráficos: Produção

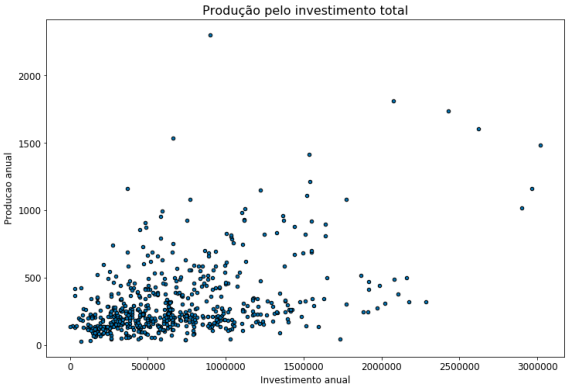
Tipos de produção por ano



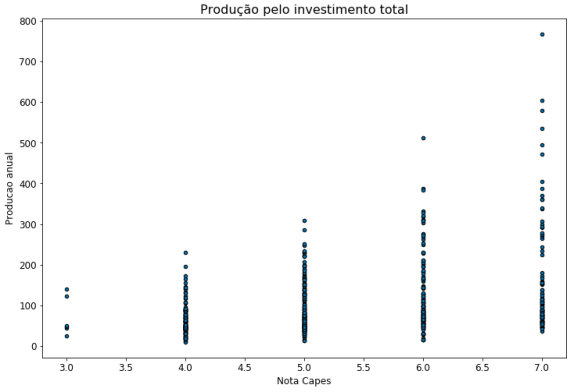
Bolsas de doutorado pelo nível CAPES



Produção pelo investimento total anual



Produção pelo investimento total de acordo com a nota CAPES

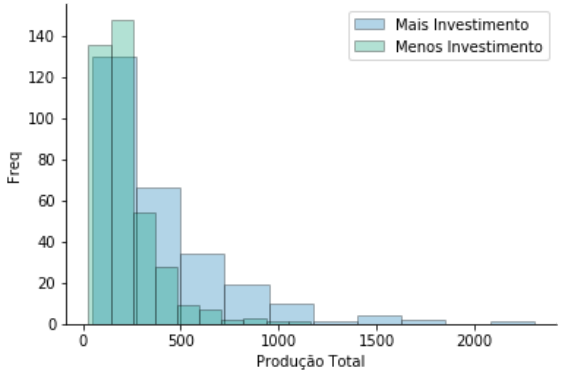


Testes de Hipótese

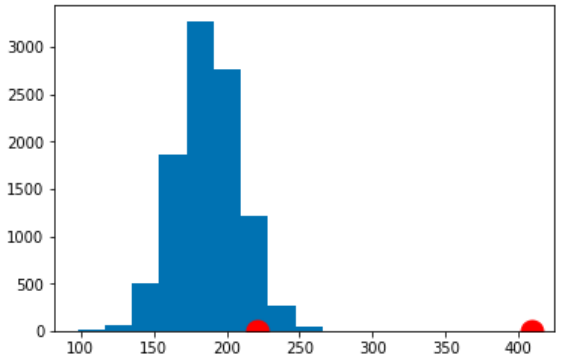
Teste 1

Hipótese nula: Na população, a distribuição das produções é a mesma para programas com investimento maior que a média e menores que a média. A diferença na amostra é devido a chance.

Hipótese alternativa: Na população, a distribuição de produção para os programas com maior investimento que a média é maior, na média, do que dos cursos com menos investimento.



Bootstrap

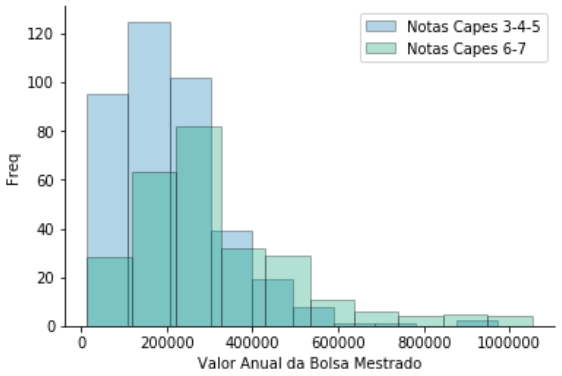


A produção dos cursos com mais investimento não está no intervalo de confiança da média da produção total de 95%. Logo **pode-se rejeitar a hipótese nula.**

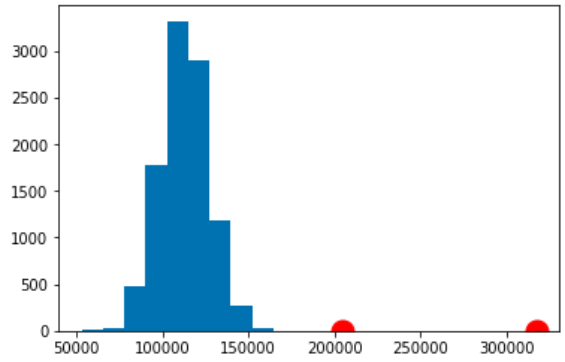
Teste 2

Hipótese nula: Na população, a distribuição do investimento é a mesma para programas com notas capes altas (maiores ou iguais a 6) e para menores ou iguais a 5. A diferença na amostra é devido a chance.

Hipótese alternativa: Na população, a distribuição do investimento para programas com notas capes altas (maiores ou iguais a 6) é maior, na média, do que para cursos com notas menores ou iguais a 5.



Bootstrap

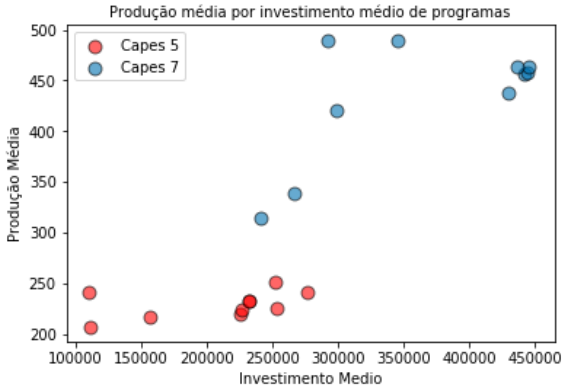


Com isso **podemos rejeitar a hipótese nula** de que na população, a distribuição do investimento é a mesma para programas com notas capes altas (maiores ou iguais a 6) e para menores ou iguais a 5. A diferença na amostra é devido a chance.

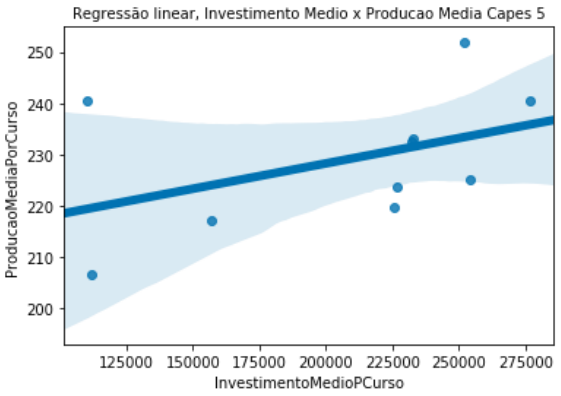
Previsão

Regressão Linear

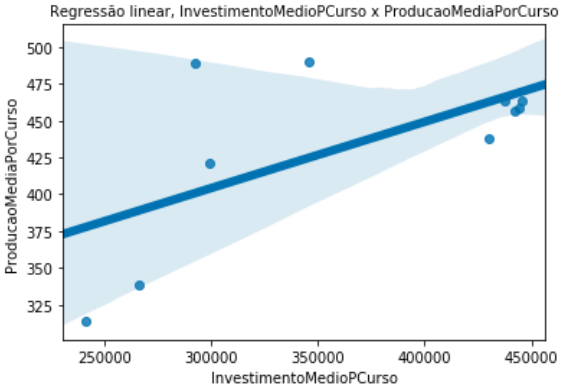
Se um curso com nível capes (3, 4 e 5) inferior recebesse o mesmo tanto que um curso CAPES 7, o quão produtivo ele seria?



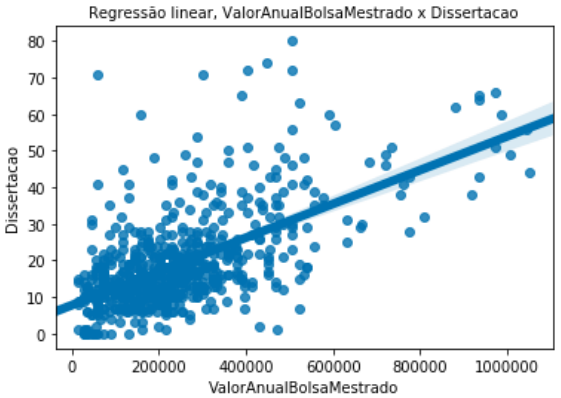
Investimento Médio x Produção Média CAPES 5



É possível ver que o investimento médio e a produção média não são tão bem explicados por um modelo linear. Vamos analisar então a produção: as dissertações.

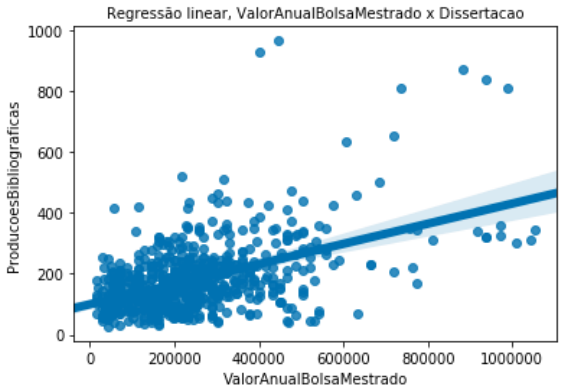


A regressão consegue explicar melhor o investimento as dissertações feitas por mestrandos.

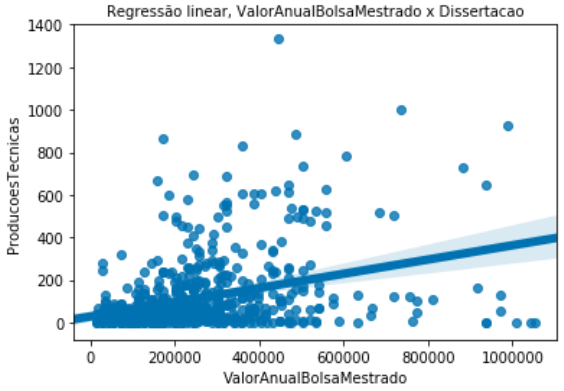


As produções bibliográficas não são bem explicadas com uma regressão linear, visto o R2 abaixo.

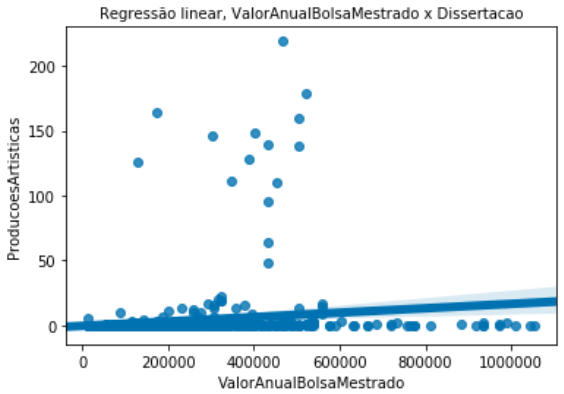
```
Coefficient: 0.00
Intercept: 96.49
Mean squared error: 10998.67
R2: 0.12
```



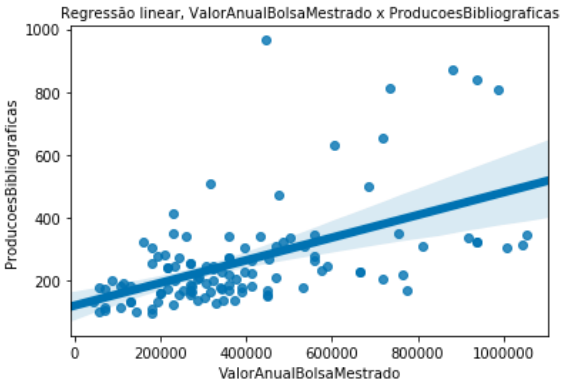
Prever as produções técnicas com o investimento através de uma regressão linear se mostra pior que do que prever utilizando a média.



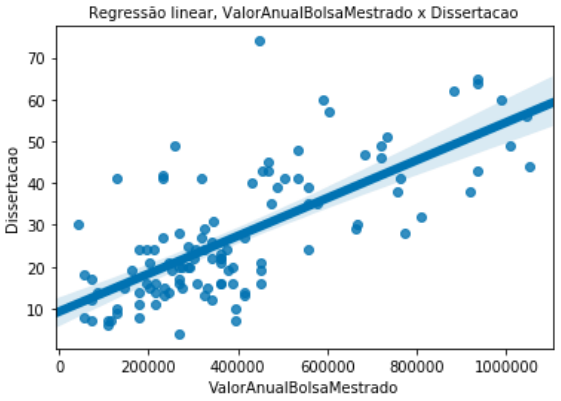
Semelhantemente o investimento nas produções artísticas não são bem previstas utilizando um modelo linear.



Já as produções bibliográficas são previstas um pouco melhor que a média.

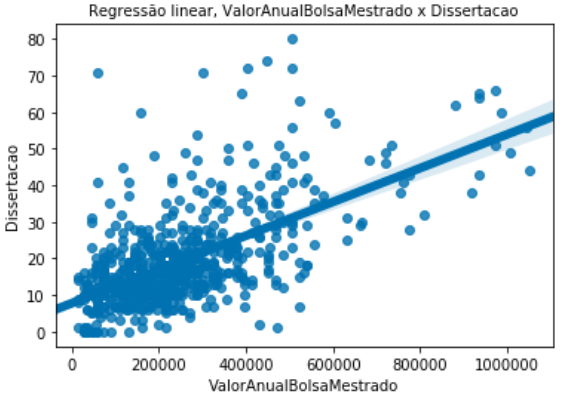


Cursos CAPES 7 tem uma relação mais forte entre investimento e as dissertações.



CAPES 5, investimento e dissertação

Cursos CAPES 5 tem uma relação mais fraca entre investimento e as dissertações.



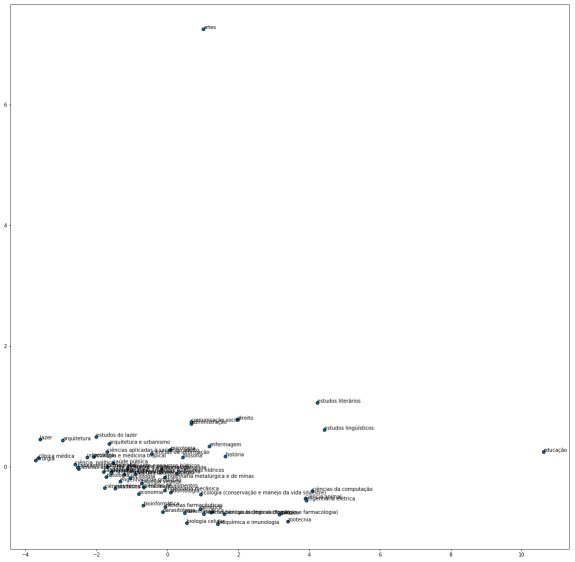
Regressão para os cursos de doutorado

As regressões para o doutorado vão ser feitas posteriormente.

Analisando PCA

Realizamos uma Análise de Componentes Principais, buscando interpretar a estrutura interna dos dados. Por ser uma técnica fundamental em reconhecimento de padrões e não ser otimizado para separabilidade de classes, não pudemos entender a natureza das relações entre o programa e a sua produção, com esse método.

PCA para os programas de pós-graduação



Classificação

O objetivo do nosso classificador é tentar estimar de forma positiva a função de classificação usada pela CAPES para definir os conceitos dos programas de pós graduação. Foi possível dessa forma construir dois classificadores, um para os programas de mestrado e outro para os de doutorado. O classificador usado foi o KNN em conjunto com uma função que busca o melhor valor de K usando cross-validation. Com o aumento da produtividade durante os anos, o dataset usado para fazer a classificação teve que ser normalizado, já que sem a normalização teríamos poucos dados e o resultado não seria melhor do que o alcançado. Os resultados abaixo mostram os dados de precisão, revocação e a matriz de confusão do nosso classificador. A qualidade do classificador poderia ser melhor com mais dados, além disso usamos apenas os dados retirados dos programas da UFMG, sendo que a avaliação leva em consideração todos os programas do Brasil.

Classificação dos cursos de mestrado

```
Precision train: 0.61
Precision test: 0.51
confusion matrix:
[[ 0  0  0  1  0]
 [ 0  9  7  1  2]
 [ 0  3 11  2  2]
 [ 0  3  0  4  2]
 [ 0  0  3  2  5]]
recall: 0.41
```

Classificação dos cursos de doutorado

```
Precision train: 0.70
Precision test: 0.42
confusion matrix:
[[0 0 0 0 1]
 [0 9 7 2 1]
 [0 3 8 7 0]
 [0 1 2 4 2]
 [0 1 2 4 3]]
recall: 0.33
```

5. Conclusões e observações finais

Críticas

Ao levantar os dados, nos deparamos com um desafio: não só de investimentos públicos é feita a ciência na UFMG, há também investimentos privados. Tentamos obter informações sobre esse tipo de investimento, mas não conseguimos devido à ausência de dados. É uma demanda da sociedade entender quanto e como foi aplicado o dinheiro investido na Ciência. Sobre esse ponto, concluímos que há uma negligência na aplicação da Lei de transparência, pois os dados públicos são de difícil compreensão e processamento, e os dados sobre investimentos privados são inexistentes.

Conclusão

Com os resultados acima, é possível perceber que, mesmo com o uso de um dataset público gerado, é possível levantar dados plausíveis e relevantes para análises. A classificação utilizada neste trabalho obteve alguns bons resultados no estudo em que ela foi proposta, mas algumas evidências de nossos resultados tiveram dificuldade de mostrar que há uma relação relevante entre investimentos e produção, devido ao fato de não termos dados suficientes para prever as notas CAPES.