Ciência de Dados Para Todos (Data Science For All) - 2019.1 - Análise da Produção Científica e Acadêmica da Universidade de São Paulo - Relatório sobre os programas de pós-graduação do Departamento de Ciência da Computação

Andre Garrido Damaceno - 15/0117531
João Marcelo Nunes Chaves - 15/0132085
Lucas Campos Jorge - 15/0154135

Introdução

A disciplina Data Science For All (Ciência de Dados para Todos) da Universidade de Brasília tem como foco integral a aplicação de ciência da dados como ferramenta eficiente de análise. O tema de estudo escolhido para realização da prática de ciência de dados foi o cenário atual da Pós-Graduação Brasileira, para que seus dados sejam processados e estudados com o objetivo de se retirar análises a respeito da qualidade, relevância, e produtividade dos programas de pós-graduação brasileiros. A fonte dos dados utilizados são dados disponíveis pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Órgão que já recolhe dados quantitativos e qualitativos sobre a pesquisa nacional, como currículos de professores e pesquisadores, além de avaliações recorrentes sobre os programas de pós-graduação a serem estudados. A interpretação e manipulação dos dados da CAPES adquiridos seguiram o modelo CRISP-DM, que guiará o projeto e será descrito ao longo do estudo. Sendo o cenário de pesquisa deste projeto a pós-graduação brasileira, foram escolhidos dois programas de pós-graduação como objetos de estudo, de onde serão extraídos seus dados. Serão estes: * Programa de pós graduação em Ciência da Computação e Matemática Aplicada pela Universidade de São Paulo USP/SC * Programa de pós graduação em Computação Aplicada pela Universidade de São Paulo USP/RP

O que é ciência?

Ciência pode ser definida como uma investigação de fenômenos desconhecidos por meio de métodos, com o objetivo de buscar a explicação, estrutura e formar uma previsão do comportamento do que é observado. Os métodos usados para a investigação devem estar em constante atualização, já que são baseados em um conjunto de princípios observados ao longo do tempo, e que não necessariamente estarão corretos para todas as situações ou que são imutáveis. Para uma obtenção fiel da análise proposta, os métodos devem ser cuidadosamente planejados. As investigações e os métodos são realizados por cientistas, por meio da produção científica, que se baseia em princípios como a projetização da ciência, racionalismo metodológico científico, empirismo científico, reprodutibilidade científica e uma comunidade científica. Como a análise de fenômenos e a consequente produção científica não possui restrições de acesso, o conhecimento é disseminado por toda a comunidade de modo que todos os interessados sejam capazes de analisar e aprimorar o conhecimento recebido. O método científico usado por cientistas é um conjunto de regras básicas que devem ser seguidas para o desenvolvimento de um padrão em que possa ser analisado o objeto de estudo de forma controlada. É composto pelos seguintes passos:

- Observação: detecção de um fenômeno.
- Investigação: análise química, física ou matemática do fenômeno.
- Problematização: identificação dos motivos e características do fenômeno.

- Hipótese: formulação de soluções e identificações de como o fenômeno funciona
- Verificação: análise final de todos os itens levantados, reprodutividade do fenômeno e obtenção de provas.

Como nem sempre todos os requisitos podem ser seguidos com total precisão, ou justamente devido a alguma falha na coleta ou formulação dos dados, a ciência não é dada como um fato absoluto. Porém, devido a seu rigor e ao avanço dos métodos e revisões, a ciência é considerada muito eficaz e precisa. Conclui-se assim que a ciência é essencial para o conhecimento humano, por estar em constante evolução, e reflete impactos em todas as áreas da sociedade através das produções científicas que exigem esforço intenso. Também sendo de suma importância na disseminação, verificação e comprovação de fenômenos e conhecimento pelo mundo.

O que é ciência no Brasil?

Ao falar em ciência no Brasil, podemos logo vinculá-la à CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Esse órgão realiza a gestão dos programas de pós graduação stricto sensu que engloba mestrados e doutorados. Por gestão entende-se acesso a informação de produções científicas, investimentos de recursos, cooperação científica internacional e outras formas de coordenação relacionadas ao desenvolvimento científico brasileiro. As produções científicas no Brasil vem, de grande parte, dos programas de pós graduação. Esses programas são formados por cursos de mestrados e doutorados oferecidos por uma instituição de ensino superior. Vale ressaltar que os cursos oferecidos devem fazer parte das áreas de conhecimento atuantes na instituição de ensino superior. Os mestrados e doutorados podem ser divididos em acadêmicos ou profissionalizantes, ou seja, uma voltada para a área acadêmica e outra voltada para a área profissional. Esses programas visam, principalmente, à formação de professorado e incentivar a pesquisa científica de alto padrão para qualificar pesquisadores e cientistas em alto nível para fazer jus ao desenvolvimento nacional do Brasil. Para que os programas de pós graduação brasileiros mantenham um nível razoável de qualidade, são necessárias avaliações periódicas realizadas pelo Sistema Nacional de Pós-Graduação Brasileira para os programas já existentes. Já para novos programas, é necessário uma avaliação da proposta inicial do programa. A CAPES possui um sistema rigoroso de avaliação para os programas nacionais de pós graduação. São realizadas avaliações a cada 4 anos e notas de 1 a 7 são atribuídas aos programas. Notas 1 e 2 são consideradas baixas e portanto não atendem o requisito mínimo para que o programa consiga prosseguir com seu andamento. A nota 3 é considerada o padrão mínimo de qualidade, ou seja, uma nota regular. Já a nota 4 é considerado um bom desempenho. Para programas que possuem apenas mestrado, a nota 5 é a maior nota. Por fim as notas 6 e 7 são consideradas altas e indicam desempenho equivalente a programas de pós graduação de excelência no exterior.

O que é CRISP-DM:

CRISP-DM (Cross-Industry Standard Process for Data Mining) é um modelo de análise de mineração de dados, feita de forma sistemática, sendo amplamente utilizada por ser flexível, podendo ser aplicada em qualquer negócio, e sua execução não ser dependente de ferramentas. As fases que compõem o CRISP-DM são as seguintes:

1. **Entendimento do negócio**: Identificação das necessidades, objetivos e tipos de soluções do negócio, a fim de transformar a realidade do negócio em um problema de data mining.
2. **Entendimento dos dados**: Reconhecer os tipos de dados disponíveis. Pode ser subdividido em cinco atividades:
    a) Análise dos dados.
    b) Realização de coleta dos dados.
    c) Descrição dos dados.
    d) Exploração dos dados: Objetivo de aprender e entender melhor a respeito.
    e) Analisar a qualidade dos dados recolhidos.

3. **Preparação dos dados**: Filtrar, a partir dos dados brutos recolhidos, visando a remoção de partes que não são necessárias, sem qualidade ou fora do contexto da mineração. Para isso, essa fase é subdividida em cinco atividades:

    a) Seleção dos dados: Identificação dos dados úteis.
    b) Limpeza dos dados.
    c) Construção dos dados: Criação de novos dados a partir da análise de outros.
    d) Integração dos dados: Unir dados de várias fontes em apenas uma.
    e) Formatação dos dados: Organização e alterações na estrutura de dados para adequação ao método de data mining escolhido.

4. **Modelagem**: É feita a construção e avaliação do modelo. O modelo segue de acordo com as quatro atividades:

    a) Seleção das técnicas de modelagem: Escolher e ajustar os parâmetros do algoritmo a ser utilizado.
    b) Realização de testes de modelagem.
    c) Construção do modelo definitivo.
    d) Avaliação do modelo e técnicas escolhidas.

5. **Avaliação**: Revisão e análise do modelo usado, organização dos dados e efeitos causados na junção de dados. Possível repetição das fases anteriores para uma maior validação.

6. **Implantação**: Inserção dos produtos desenvolvidos para uso, com um monitoramento regular, sendo feitas adaptações e ajustes quando necessários. Essa fase marca o fim da produção do projeto.

As metodologias propostas pelo CRISP-DM estão apresentados no diagrama abaixo:

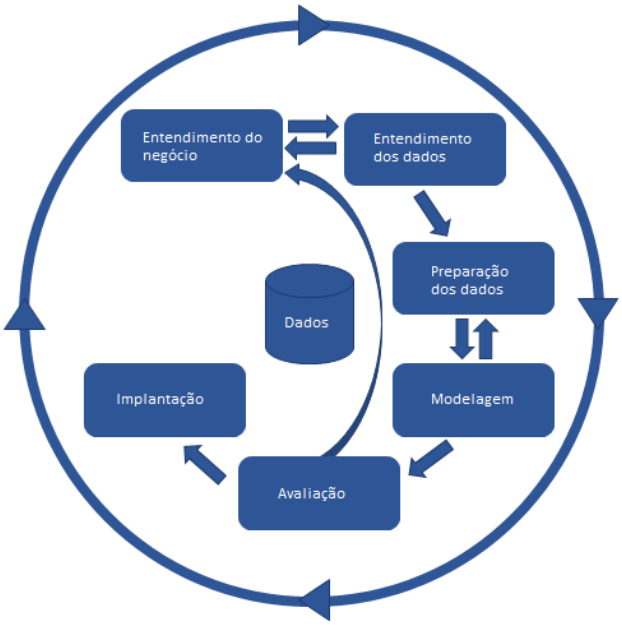

Figure 1: Diagrama CRISP-DM

Contexto dos programas de pós-graduação

Programa de pós graduação em Ciência da Computação e Matemática Aplicada pela Universidade de São Paulo USP/São Carlos

O programa de pós graduação em Ciência da Computação e Matemática Aplicada está organizado em 5 linhas de pesquisa sendo elas: * Computação Gráfica, Imagens e Visualização * Engenharia de Software e Sistemas de Informação/Sistemas Web e Multimídia Interativos * Inteligência Computacional * Sistemas Distribuídos e Programação Concorrente/Sistemas Embarcados, Evoluídos e Robóticos * Mecânica dos Fluidos Computacional/Otimização/Modelos Estocásticos

As disciplinas ofertadas para os discentes englobam as principais áreas da Ciência da Computação. Já seu corpo docente é de alto nível, contendo professores formados em ótimas universidades no Brasil e no exterior.

O programa conta como modalidades o mestrado e o doutorado acadêmico, tendo em vista formar pesquisadores e professores na área. O curso é muito bem avaliado e recebeu a nota 7 (máxima) na avaliação trienal da CAPES. Para tanto, o curso foi avaliado como tendo: uma excelente qualificação e inserção internacional dos docentes; dissertações e teses defendidas de ótima qualidade; Ótima produção bibliográfica; e uma gama de atividades de cooperação internacional

Discentes	452
Docentes	71
Disciplinas	292
Financiadores	5
Linhas de Pesquisa	5
Projetos de Pesquisa	88

Programa de pós graduação em Computação Aplicada pela Universidade de São Paulo USP/RP

O programa de pós graduação em computação aplicada possui como modalidade, apenas, o mestrado acadêmico e contempla duas linhas de pesquisa, sendo elas: Computação aplicada à Biociências e Sistemas Computacionais Complexos.

As disciplinas ofertadas pelo programa abrangem as três áreas da Ciência da Computação, sendo elas:

- Metodologia e Técnicas de Computação
- Teoria da Computação e Análise de Algoritmos e Complexidade da Computação
- Sistemas de Computação

O programa, apesar de recente, conta com um corpo docente experiente onde cada docente possui, pelo menos uma, orientação de mestrado e a maioria possui pelo menos uma orientação de doutorado concluída. Apesar disso, o programa conta com poucos docentes com dedicação exclusiva para o programa, o que é considerado um ponto em que o programa deixa a desejar. Na última avaliação trienal CAPES, o programa recebeu nota 3. Tal nota foi justificada principalmente pela falta de docentes com dedicação exclusiva, também foi ponderado a falta de disciplinas que fazem com que o discente não garanta aproveitamento em cada uma das áreas do núcleo básico da ciência da computação. Por ser um programa de pós graduação muito novo, ainda não há publicações discentes e portanto não existe uma maneira de qualificar as produções científicas.

Discentes	23
Docentes	15
Disciplinas	21
Financiadores	5
Linhas de Pesquisa	2
Projetos de Pesquisa	15

Análise dos dados coletados

Para a coleta de dados dos programas de graduação citados acima, utilizamos a plataforma elattes (http://unb.elattes.com.br). Por meio dessa plataforma, obtivemos os dados de 2014 até 2018 das publicações científicas, dos docentes afiliados e sobre as orientações de cada programa de pós-graduação. Com auxilio da linguagem R, utilizamos scripts que nos permitiram interpretar os dados coletados.

Computação Aplicada

```
library(tidyverse); library(jsonlite); library(listviewer);

#Ler o arquivo profile
profile <- jsonlite::fromJSON("comp_aplicada/profile.json")
public <- jsonlite::fromJSON("comp_aplicada/publication.json")
advise <- jsonlite::fromJSON("comp_aplicada/advise.json")

#Arquivo com funcionalidades que transformam o arquivo formato list em DataFrames
source("scripts/elattes.ls2df.R")

concluidas <- advise %>% names() %>% grepl(pattern = "CONCLUIDA")

for (i in names(advise[concluidas])){
  print(i)
  print(advise[[i]] %>%
    sapply(function(x) length(x$natureza)))
}

# Analise Profile

orient.posdoutorado.df <- ori.ls2df(advise, 6) #pos-Doutorado concluído
orient.doutorado.df <- ori.ls2df(advise, 7) #Doutorado concluído
orient.mestrado.df <- ori.ls2df(advise, 8) #Mestrado concluído

orient.posdoutorado_p.df <- ori.ls2df(advise, 1) #pos-Doutorado concluído
orient.doutorado_p.df <- ori.ls2df(advise, 2) #Doutorado concluído
orient.mestrado_p.df <- ori.ls2df(advise, 3) #Mestrado concluído

orient.df <- rbind(rbind(orient.posdoutorado.df, orient.doutorado.df), orient.mestrado.df)
orient_p.df <- rbind(rbind(orient.posdoutorado_p.df, orient.doutorado_p.df), orient.mestrado_p.df)

orient_sum <- group_by(orient.df, ano, natureza) %>%
                summarise(n = n())
orient_p_sum <- group_by(orient_p.df, ano, natureza) %>%
  summarise(n = n())

ggplot(orient_sum, aes(x = ano, y = n, group = natureza, color = natureza)) +
  geom_line(alpha = 0.6) +
  geom_point(alpha = 0.6)

ggplot(orient_p_sum, aes(x = ano, y = n, group = natureza, color = natureza)) +
    geom_line(aes(alpha = 0.3)) +
    geom_point(aes(alpha = 0.3))
```

```
# extrai producao bibliografica de todos os professores
pr.df.pub <- extrai.producoes(profile)

#Eventos por país
pr.df.pub %>%
  filter(tipo_producao == 'EVENTO') %>%
  group_by(pais_do_evento) %>% summarize(n = n()) %>%
  ggplot() + geom_bar(mapping = aes(x = factor(pais_do_evento), y = n), fill = '#1287F6', stat = "ident

# Número de produções em eventos por ano
pr.df.pub %>%
  filter(tipo_producao == 'EVENTO') %>%
  group_by(ano_do_trabalho) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = ano_do_trabalho, y = n), fill = '#1287F6', stat = "identity")


# Número de produções em periodicos por ano
  pr.df.pub %>%
  filter(tipo_producao == 'PERIODICO') %>%
  group_by(ano) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = ano, y = n), fill = '#1287F6', stat = "identity")

  pr.df.areas <- extrai.areas.atuacao(profile)

# Area de atuação grande area:
  pr.df.areas %>%
  group_by(grande_area) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = grande_area, y = n), fill = '#1287F6', stat = "identity") + coord_flip()

# Area de atuação area:
pr.df.areas %>%
group_by(area) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = area, y = n), fill = '#1287F6', stat = "identity") + coord_flip()

# Area de atuação especialidade:
  pr.df.areas %>%
  filter(especialidade != '') %>%
  group_by(especialidade) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = especialidade, y = n), fill = '#1287F6', stat = "identity") + coord_flip()

# Financiamento de Orientações concluidas de pós-graduação:
orient.df %>%
  filter(agencia_financiadora != '') %>%
  group_by(agencia_financiadora) %>% summarize(n = n()) %>%
  ggplot() + geom_bar(mapping = aes(x = agencia_financiadora, y = n), fill = '#1287F6', stat = "identity

# Financiamento de Orientações em andamento de pós-graduação do ano de 2018:
orient_p.df %>%
```

```
  filter(agencia_financiadora != '' & ano == 2018) %>%
  group_by(agencia_financiadora) %>% summarize(n = n()) %>%
  ggplot() + geom_bar(mapping = aes(x = agencia_financiadora, y = n), fill = '#1287F6', stat = "identity
```

Ciência da Computação e Matemática Aplicada

```
library(tidyverse); library(jsonlite); library(listviewer);

#Ler o arquivo profile
profile <- jsonlite::fromJSON("cic/profile.json")
public <- jsonlite::fromJSON("cic/publication.json")
advise <- jsonlite::fromJSON("cic/advise.json")

#Arquivo com funcionalidades que transformam o arquivo formato list em DataFrames
source("scripts/elattes.ls2df.R")

concluidas <- advise %>% names() %>% grepl(pattern = "CONCLUIDA")

for (i in names(advise[concluidas])){
  print(i)
  print(advise[[i]] %>%
    sapply(function(x) length(x$natureza)))
}

# Analise Profile

orient.posdoutorado.df <- ori.ls2df(advise, 6) #pos-Doutorado concluído
orient.doutorado.df <- ori.ls2df(advise, 7) #Doutorado concluído
orient.mestrado.df <- ori.ls2df(advise, 8) #Mestrado concluído

orient.posdoutorado_p.df <- ori.ls2df(advise, 1) #pos-Doutorado concluído
orient.doutorado_p.df <- ori.ls2df(advise, 2) #Doutorado concluído
orient.mestrado_p.df <- ori.ls2df(advise, 3) #Mestrado concluído

orient.df <- rbind(rbind(orient.posdoutorado.df, orient.doutorado.df), orient.mestrado.df)
orient_p.df <- rbind(rbind(orient.posdoutorado_p.df, orient.doutorado_p.df), orient.mestrado_p.df)

orient_sum <- group_by(orient.df, ano, natureza) %>%
                summarise(n = n())
orient_p_sum <- group_by(orient_p.df, ano, natureza) %>%
  summarise(n = n())

ggplot(orient_sum, aes(x = ano, y = n, group = natureza, color = natureza)) +
  geom_line(alpha = 0.6) +
  geom_point(alpha = 0.6)

ggplot(orient_p_sum, aes(x = ano, y = n, group = natureza, color = natureza)) +
    geom_line(aes(alpha = 0.3)) +
    geom_point(aes(alpha = 0.3))

# extrai producao bibliografica de todos os professores
```

```
pr.df.pub <- extrai.producoes(profile)

#Eventos por país
pr.df.pub %>%
  filter(tipo_producao == 'EVENTO') %>%
  group_by(pais_do_evento) %>% summarize(n = n()) %>%
  ggplot() + geom_bar(mapping = aes(x = factor(pais_do_evento), y = n), fill = '#1287F6', stat = "ident

# Número de produções em eventos por ano
pr.df.pub %>%
  filter(tipo_producao == 'EVENTO') %>%
  group_by(ano_do_trabalho) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = ano_do_trabalho, y = n), fill = '#1287F6', stat = "identity")


# Número de produções em periodicos por ano
  pr.df.pub %>%
  filter(tipo_producao == 'PERIODICO') %>%
  group_by(ano) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = ano, y = n), fill = '#1287F6', stat = "identity")

  pr.df.areas <- extrai.areas.atuacao(profile)

# Area de atuação grande area:
  pr.df.areas %>%
  group_by(grande_area) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = grande_area, y = n), fill = '#1287F6', stat = "identity") + coord_flip()

# Area de atuação area:
pr.df.areas %>%
group_by(area) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = area, y = n), fill = '#1287F6', stat = "identity") + coord_flip()

# Area de atuação especialidade:
  pr.df.areas %>%
  filter(especialidade != '') %>%
  group_by(especialidade) %>% summarize(n = n()) %>%
  ggplot() +
  geom_bar(mapping = aes(x = especialidade, y = n), fill = '#1287F6', stat = "identity") + coord_flip()

# Financiamento de Orientações concluidas de pós-graduação:
orient.df %>%
  filter(agencia_financiadora != '') %>%
  group_by(agencia_financiadora) %>% summarize(n = n()) %>%
  ggplot() + geom_bar(mapping = aes(x = agencia_financiadora, y = n), fill = '#1287F6', stat = "identity

# Financiamento de Orientações em andamento de pós-graduação do ano de 2018:
orient_p.df %>%
  filter(agencia_financiadora != '' & ano == 2018) %>%
  group_by(agencia_financiadora) %>% summarize(n = n()) %>%
```

```
ggplot() + geom_bar(mapping = aes(x = agencia_financiadora, y = n), fill = '#1287F6', stat = "identity
```

Referências

[1] Jorge Henrique Cabral Fernandes, Ricardo Barros Sampaio. Unb Aprender, 2019. Sobre a Ciência e sua Avaliação. Disponível em https://aprender.ead.unb.br/pluginfile.php/474549/mod_resource/content/2/SobreCiencia.pdf. Acesso em: 13/04/2019.

[2] IBM. IBM, 2012. CRISP-DM Help Overview Disponível em https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm. Acesso em: 13/04/2019.

[3] Wikipedia. Wikipedia, 2019. Cross Industry Standard Process for Data Mining Disponível em: https://pt.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining. Acesso em: 13/04/2019.

[4] ICMC-USP. ICMS USP, 2019 Disponível em: https://icmc.usp.br/pos-graduacao/ppgccmc. Acesso em: 13/04/2019.

[5] Avaliação Quadrienal. Ficha de Avaliação do Programa CIÊNCIAS DA COMPUTAÇÃO E MATEMÁTICA COMPUTACIONAL (33002045004P1) UNIVERSIDADE DE SÃO PAULO (USP) Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/avaliacao/consultaFichaAvaliacao.xhtml. Acesso em: 13/04/2019.

[6] Avaliação Quadrienal. Ficha de Avaliação do Programa COMPUTAÇÃO APLICADA (33002029052P5) UNIVERSIDADE DE SÃO PAULO (USP) Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/avaliacao/consultaFichaAvaliacao.xhtml. Acesso em: 13/04/2019.

[7] Brasil Escola. Método científico Disponível em: https://brasilescola.uol.com.br/quimica/metodo-cientifico.htm. Acesso em: 22/05/2019.