Teoria e Prática em Administração, v. 11, n. 1.
Moraes, Valdevino & Nascimento
Seção: caso para ensino
DOI: https://doi.org/10.22478/ufpb.2238-104X.2021v11n1.51551

Serenata de Amor: um Doce Não Tão Saboroso

Love Serenade: a Sweet Not So Tasty

Thiago Moraes
Universidade Estadual do Piauí – UESPI – Brasil
thgmoraes@gmail.com
ORCID: 0000-0001-9729-4858

Antonio Messias Valdevino
Universidade Federal do Cariri– UFCA – Brasil
messiasurca@gmail.com
ORCID: 0000-0001-7096-7071

Anderson Lopes Nascimento
Universidade Federal do Piauí – UFPI – Brasil
adm.anderson@hotmail.com
ORCID: 0000-0003-2538-9000

Submetido em 30/03/2020; **Aprovado** em 05/05/2020.

Resumo
Objetivo: Este caso para ensino discute as aplicações de uma iniciativa cívica de monitoramento de gastos públicos dos deputados federais do Brasil. **Fontes de Dados:** É um projeto de *Data Science* aberto que utiliza as mesmas tecnologias de inteligência artificial e de *Machine learning* empregadas por gigantes como Google, Microsoft e Netflix para fiscalizar o uso das verbas indenizatórias concedidas aos parlamentares e compartilhar as informações de forma acessível a qualquer pessoa, de modo que o cidadão com um mínimo de conhecimento possa fiscalizar e auditar tais gastos. **Contexto:** A Serenata criou a Rosie, uma inteligência artificial capaz de analisar os gastos reembolsados pela Cota para Exercício da Atividade Parlamentar (CEAP), de deputados federais e senadores, feitos em exercício de sua função, identificando suspeitas e incentivando a população a questioná-los. A Serenata criou também o Jarbas, um dashboard que facilita a visualização de todas as informações monitoradas pela Rosie. **Aplicação**: Em disciplinas de cursos tecnólogos, graduação e pós-graduação, *lato sensu* e *strictu sensu* de Administração, Sistemas de Informação e seus afins, em temas relacionados a *Data Science, Machine Learning* e Dados Abertos.
Palavras-chave: Aprendizagem de Máquina, Ciência de Dados, Inteligência Artificial, Verba Indenizatória.

Abstract
Objective: This teaching case seeks to discuss the applications of a civic initiative to monitor public spending by federal deputies in Brazil. **Data Source**: It is an open Data Science project that uses the same artificial intelligence and Machine learning technologies used by giants like Google, Microsoft and Netflix to monitor the use of indemnity funds (Quotas for the Exercise of Parliamentary Activity) granted to parliamentarians and share information in an accessible way to anyone, so that the citizen with a minimum of knowledge can inspect and audit such expenses. **Context**: The Serenata created Rosie, an artificial intelligence capable of analyzing the expenses reimbursed by the Quota for Exercising Parliamentary Activity of federal deputies and senators, made in the exercise of their function, identifying suspicions and encouraging the population to question them and with her, Jarbas, a dashboard that makes it easy to view all the information monitored by Rosie. **Application**: disciplines of technologist courses, undergraduate and graduate courses, lato sensu and strictu sensu of Administration, Information Systems and the like, in subjects related to Data Science, Machine Learning and Open Data.
Keywords: Machine Learning, Data Science, Artificial Intelligence, Compensation Allowance.

Introdução

- Oi, eu sou um robô. Meu nome é Rosie e estou aqui para representar a Operação Serenata de Amor. Foi a Serenata quem me criou e me deu um trabalho, mais que um trabalho, uma missão. Eu analiso os gastos reembolsados pela Cota para Exercício da Atividade Parlamentar (CEAP), conhecidas como verbas indenizatórias, de deputados e senadores para identificar ações e gastos suspeitos.

Imagine a seguinte situação: uma pessoa fazer treze refeições no mesmo dia ou consumir bebida alcoólica em expediente ou numa viagem a trabalho. Pode parecer anormal, mas acrescente o peso de que quem efetuou esse gasto foi um deputado ou um senador que pagou essas depesas com dinheiro público.

Samuel, que é um empreendedor de tecnologias sociais que utiliza dados abertos governamentais para desenvolver soluções para a sociedade por meio de *Data Science*, diante de casos de gastos discrepantes dos membros do Congresso Nacional, decidiu explorar a forma com que os deputados gastam o dinheiro público e parou para refletir sobre isso:

- Parece que há algo de errado aí, não é mesmo? Mas, será que seria apenas essa situação? O que mais de errado pode haver nas contas prestadas pelos parlamentares e congressistas?

E, se aprofundando no trabalho da Rosie e na Operação Serenata de Amor, Samuel buscou responder suas dúvidas. E a primeira delas foi:

- De onde surgiu esse nome, Serenata de Amor? Por quê? Com que objetivo essa iniciativa nasceu? Que perguntas ainda não foram respondidas? Que perguntas ainda não foram feitas aos dados?

A Operação Serenata de Amor

Serenata de Amor (https://serenata.ai/) é o nome dado a uma operação que tem a finalidade de fiscalizar gastos públicos. O nome foi dado em alusão a um escândalo ocorrido nos anos 90, na Suécia, o "Caso Toblerone", que consistiu em uma investigação de compras de chocolate daquela marca. Nessa investigação, descobriu-se que a proponente mais cotada ao cargo de primeira-ministra do país europeu teria utilizado verba pública para cobrir gastos pessoais.

Criada em 07 de setembro de 2016, juntando controle social e tecnologia, a Serenata é uma ferramenta capaz de ajudar a auditoria e fiscalização das contas públicas. A operação se dá por meio de aprendizagem de máquina, Rosie, a robô e personagem principal da narrativa. Rosie todo dia aprende coisas novas e expande o projeto para levar mais informações para a população a partir da transformação de dados públicos em conteúdo acessível.

A Serenata do amor é um projeto que representa uma tecnologia de inovação cívica

Tecnologias cívicas para soluções sociais. Isso foi tudo que Samuel sempre buscou para sua *Startup*, a qual, coincidentemente, você acaba de ser contratado como cientista de dados júnior.

- Pessoal, parem tudo o que estão fazendo e olhem para mim. Vamos explorar o projeto Serenata de Amor e levantar mais questionamentos dentro do que ele nos pode oferecer.

Assim, Samuel desafiou sua equipe. Ao que seu analista de dados retrucou:

- Samuel, por onde começamos?

- Primeiro, vamos entender o propósito da iniciativa. Em seguida, vamos aos dados que eles usam como fontes. Quero também que vocês investiguem cada deputado e senador que já foi mencionado no Twitter da operação, seu reduto eleitoral, tempo de mandato, quantidade de projetos propostos e aprovados, tendência de votação, financiamento da campanha etc.

A visão de Samuel faz sentido porque a operação Serenata de Amor começou como uma ideia de ciência de dados, que busca usar a tecnologia para navegar pela política e os dados que ela fornece. Isso se dá, principalmente, por meio da Lei de Acesso à Informação, Lei 12.527/2011, (LAI), que trata a informação como direito humano fundamental, com base no artigo 19 da Declaração Universal dos Direitos Humanos, adotada pela Assembleia Geral da Organização das Nações Unidas (ONU).

A primeira versão do projeto era focada em analisar os dados das verbas indenizatórias da Câmara dos Deputados, observando gastos suspeitos, de uma forma que manualmente teria um custo muito alto, de tempo e de processamento. Aqui, a inteligência artificial denominada de Rosie foi idealizada e cons-

truída. É um robô que faz uma varredura nos dados das verbas indenizatórias da Câmara Federal, analisando um volume muito grande deles e originando as técnicas de *machine learning* usadas na operação.

O projeto buscou financiamento coletivo por meio do *crowdfunding*. Inicialmente, na plataforma Catarse e, em seguida, no APOIA.SE para custear a equipe, na qual a Rosie foi desenvolvida, alcançando 629 denúncias da Câmara dos Deputados, em 2016. A plataforma foi capaz de analisar mais de 3 milhões de notas, levantando cerca de 8.000 casos suspeitos em gastos com verbas públicas. A comunidade que suporta o trabalho do time se beneficia de repositórios de código aberto, licenciados para a colaboração, o que levou as duas principais cientistas de dados do projeto o apresentarem na CivicTechFest, em Taipei, obtendo várias menções inclusive na imprensa internacional.

Há uma parte técnica e de educação política do projeto

O *dataset* de reembolsos coletados pela Rosie já considera os dados da empresa, dizendo quem é o seu sócio, por exemplo. Não é um volume de dados considerado como *Big Data*, contudo, não há ainda a possibilidade de manipulá-los em plataformas como o Excel.

As principais dificuldades:

- A qualidade dos arquivos é uma barreira. O *encoding* do arquivo, diferente do declarado, dificulta a fase de Extração, Tratamento e Carregamento dos Dados (ETL). Os metadados não facilitam a leitura (UTF-8, por exemplo); a granularidade e o *encoding* variam; e não há uma diretriz padrão para liberação de dados. Afirmou, apreensivo, um dos programadores da *startup* de Samuel, ao que ele responde:

- Então, vamos ao *Github* da operação, que apesar de estar em inglês para se adequar à comunidade *open source* e se ajustar à possibilidade de algum outro país se interessar pela iniciativa e replicar assim como nós, é, de fato, um ato de incentivo à transparência e combate à corrupção. E vamos encampar essa briga.

Samuel descobriu, a duras penas, que a Câmara dos Deputados disponibilizava os dados em grande volume e formato aberto, mas pouco acessível. Motivo pelo qual a Operação Serenata de Amor passou a usar um formato de dado que reduz o tamanho e exige uma menor capacidade de processamento. Isso exige menos espaço de armazenamento, reduzindo custos relacionados à análise e processamento, o que se configura uma vantagem para uma *startup* como a de Samuel, ainda nova e com recursos financeiros limitados.

A tecnologia

A equipe de Samuel usa o Jupyter Notebook para manipular os dados. É um ambiente computacional *web* para a internet, feito para criação de documentos compartilhados (*Notebooks*). Assim, são feitas as Análises Exploratórias de Dados (AED's), é o mesmo procedimento realizado pela Operação. Nesse processo, são testadas hipóteses nas quais os classificadores recebem os dados, olham os reembolsos e os classificam como suspeito ou não (papel da Rosie). Desta forma, quando o *notebook* se mostra efetivo, ele é reescrito na Rosie.

Samuel compreendeu que quando um teste de hipótese é levantado e terminado, ele pode se tornar um classificador a ser implementado na Rosie. A detecção de um *outlier* de refeição é um exemplo. Definidos os passos para se identificar uma suspeita, classifica-se o que é ou não gasto suspeito e chama-se a Rosie para testar a hipótese. Ela é um código python de processos *Fit*, *Predict* e *Transform*, onde o csv (arquivos *comma-separated values*) gerado leva à suspeição ou não do gasto. Daí, o projeto Jarbas se faz importante. É um site no qual são disponibilizados os dados da câmara, os resultados da Rosie, os *links* para a nota fiscal escaneada e os endereços do *Google Street View* dos estabelecimentos onde o gasto foi realizado.

- Pessoal, deixa eu explicar! Afirmou Samuel, frente a um quadro branco na única sala que a startup dele tem, um grande galpão, sem divisórias e cheia de fios, computadores, banheiro e uma minicozinha.

- A Rosie atualmente pede ajuda à população, pelo twitter ela avisa que achou um gasto suspeito, marca o deputado e direciona o *link* para o Jarbas. Qualquer um de nós, ou seja, qualquer cidadão, pode abrir o *link* do gasto, onde aparece o arquivo e formato PDF. Lá, está a foto da nota fiscal do gasto, *touché*!

- O Jarbas – continua Samuel - é, basicamente, a forma visual de apresentar o gasto suspeito, é um painel para acesso à informação sobre o parlamentar e o uso do recurso público. Como algoritmo, a Rosie tem suas particularidades e usa outras como suporte. O *K-Means* é usado sobre preços para

agrupá-los por características.

Theodora, uma estagiária, do oitavo período de Administração, levanta a mão e pergunta:

-Então, Samuel, a Rosie é um projeto que trabalha com dados abertos e permite que pessoas, mesmo sem habilidade técnicas, possam colaborar?

- Exatamente! Isso é o que evidencia o propósito colaborativo e cívico da ideia. Responde Samuel. É aqui que nós entramos! Mas todos precisam entender tudo sobre as cotas para exercício da atividade parlamentar, certo?

O que é a CEAP?

A Cota para o Exercício da Atividade Parlamentar (CEAP) é um valor mensal a ser pago aos Deputados Federais que objetiva custear seus gastos exclusivamente vinculados ao exercício da atividade parlamentar (Câmara dos Deputados, 2020).

O seu uso é regulado pelo Ato da Mesa nº 43 de 2009, que discrimina os gastos classificados como indenizatórios, como passagens aéreas; telefonia; serviços postais; fornecimento de alimentação ao parlamentar; hospedagem; combustíveis e lubrificantes etc., excetuando-se os 120 dias anteriores às eleições; participação do parlamentar em cursos, palestras, seminários, simpósios, congressos ou eventos congêneres; e a complementação do auxílio-moradia. (Câmara dos Deputados, 2020).

O valor da CEAP é determinado de acordo com a unidade da federação a que pertence o parlamentar. É a partir da CEAP que a operação começa a monitorar, por meio da Rosie, os gastos por partido e os gastos de cada deputado. Assim, em caso de suspeita, o parlamentar é procurado, informado e disponibilizado a ele o espaço para resposta e justificativa do gasto. Caso a resposta seja compreendida como caso de corrupção ou desvio de finalidade, a Operação direciona uma denúncia ao Ministério Público e as informações sobre os gastos são publicadas nas redes sociais da Serenata do Amor.

Um Problema a mais?

Um dos problemas enfrentados pela Serenata foi o tecnológico, como o formato dos arquivos. Estes possuíam baixo nível de acessibilidade. Outro problema diz respeito ao próprio ato que rege a CEAP, que não proíbe expressamente o uso da verba para gastos com bebidas alcóolicas, por exemplo. No entanto, tem-se registro de reembolso de valores relativos a este item de consumo não contestados à época, 2015. Aqui, o *link* do portal da Câmara dos Deputados (https://www.camara.leg.br/cota-parlamentar/documentos/publ/2880/2015/5660757.pdf) para acesso ao comprovante de consumo de uma cerveja da marca Samuel Adams (https://www.samueladams.com/), no restaurante do Gordon Ramsay (https://www.gordonramsay.com/), conhecido em um famoso programa de TV. Outro destaque que chama a atenção é o do parlamentar que efetuou pagamentos a uma empresa criada por ele mesmo e registrou o comprovante para reembolso, acessível no link a seguir (https://www.camara.leg.br/cota-parlamentar/documentos/publ/705/2015/5621548.pdf). Neste caso, a empresa estava inativa na Receita Federal desde 2014 e os comprovantes datavam de 2015.

O que queria então Samuel de sua equipe a qual você faz parte?

- Se uma AED pode extrair informações interessantes como o valor de reembolso de uma refeição no valor de R$ 6.205,00 ou o de 13 refeições realizadas no mesmo dia, por um mesmo deputado, podemos supor que, ao analisar cada caso exige um custo e tecnologia para combater à corrupção usando conhecimento em *Data Science*. Isso, nós temos aqui de sobra.

- Então, pessoal – Samuel levanta da cadeira e aponta para o quadro branco - desta forma, eu desafio vocês, a até o fim do dia, apresentar neste quadro dilemas ainda não levantados pela operação que podem contribuir para o combate a corrupção usando nosso *know-how* em tecnologia.

Às 19h, do mesmo dia, Samuel, o último a sair da firma, levanta da sua baia e passa pela parede onde fica aquele quadro branco. Surpreso e contente com o que virá no dia seguinte, saca seu smartphone e tira uma foto do que foi escrito ali para que também possa refletir sobre elas quando chegar em casa.

1) Que relações existem entre os gastos de deputados federais com verba indenizatória e os CNPJ das empresas onde os gastos foram feitos?
2) Considerando endereços e localizações, a agenda oficial do político, suas votações, os nomes dos sócios das empresas; e data de abertura e baixa da empresa como achar dados que desviam dos padrões? O que dizem os padrões achados?

3) Qual a frequência de aparecimento de nomes de familiares nas empresas e se há contribuições dessas empresas nos financiamentos de campanha dos partidos ou do próprio deputado? Existe relação entre o financiamento da campanha do parlamentar e sua votação?

NOTAS DE ENSINO

Este caso para ensino trata de um projeto de *Data Science* aberto, criado para fiscalizar o uso das verbas indenizatórias (Cotas para o Exercício da Atividade Parlamentar) de deputados federais e senadores, feitas no exercício de suas funções, identificando suspeitas e incentivando a população a questioná-las e compartilhar as informações de forma acessível a qualquer pessoa.

Fontes dos Dados

Todas as informações apresentadas no enredo do caso são de reportagens sobre a Serenata de Amor, de entrevistas com membros da equipe de desenvolvimento e de análise de dados. A página oficial da Operação e as redes sociais serviram de fonte para a construção deste caso, bem como o *twitter* da Rosie (https://twitter.com/RosieDaSerenata).

A construção do caso considerou fontes de dados primárias, nas quais o professor pode, junto com sua turma em laboratório, acessar os dados da Câmara Federal no link_próprio (https://dadosabertos.camara.leg.br/) ou no Github da Serenata (https://github.com/okfn-brasil/serenata-de-amor). Ainda no Github da Serenata, é possível encontrar os códigos referentes a operação, escritos em linguagem Python. Para tanto, recomenda-se que na aplicação deste caso, os alunos tenham uma noção introdutória da linguagem de programação, por isso deve-se obedecer às diretrizes de aplicação a seguir. Uma alternativa ao uso do Python é o Excel, por meio de *Add-ins* que facilitam as análises requisitadas no caso.

Objetivos educacionais e Diretrizes de aplicação do caso para ensino

O objetivo educacional do caso é discutir o conhecimento gerado por um projeto de *Data Science* de tecnologia cívica de combate à corrupção por meio da inteligência artificial, bem como refletir sobre seu processo de construção, tendo como fonte principal os dados abertos da Câmara Federal de Deputados. Assim, recomenda-se que o aluno seja estimulado a identificar nas bases de dados abertas, e por meio de uma Análise Exploratória de Dados, as informações que possam auxiliar sua decisão e o suporte na identificação de práticas de corrupção, mais especificamente de desvio de finalidade no uso da CEAP. Essas bases estão disponíveis nos *links* ativados no texto e na seção de *links* úteis.

Nesse sentido, deve-se considerar que os dados abertos governamentais devem ser (TCU, 2015):

- **Completos**: todos os dados públicos estão disponíveis. Não sujeitos a limitações válidas de privacidade, segurança ou controle de acesso.
- **Primários:** apresentados tais como os coletados na fonte, com o maior nível de granularidade e sem agregação ou modificação.
- **Atuais:** disponibilizados tão rapidamente quanto necessária à preservação do seu valor.
- **Acessíveis:** disponibilizados para o maior alcance possível de usuários e para o maior conjunto possível de finalidades.
- **Compreensíveis por máquinas:** os dados são razoavelmente estruturados de modo a possibilitar processamento automatizado.
- **Não discriminatórios:** os dados são disponíveis para todos, sem exigência de requerimento ou cadastro.
- **Não proprietários:** os dados são disponíveis em formato sobre o qual nenhuma entidade detenha controle exclusivo;
- **Livres de licenças:** os dados não estão sujeitos a nenhuma restrição de direito autoral, patente, propriedade intelectual ou segredo industrial.

Para o caso, as etapas a seguir devem ser adotadas pelo docente:

A leitura prévia individual deve ser fortemente recomendada, e como parte da atividade, o discente deve destacar palavras e termos desconhecidos ou que lhe chamem atenção e pesquisar sobre eles. Esta etapa deve ser realizada em casa, confeccionando um seguinte esboço de Data Science:

1) Identificação da demanda
2) Compreensão do problema
3) Procedimentos de coleta de dados
4) Processamento e exploração de dados/Análise de dados
5) Comunicação de resultados

Este caso é indicado para aplicação em turmas de, no mínimo 20 alunos e, em laboratório de informática, com conexão de internet ativa e estável. Em caso de indisponibilidade de um laboratório, sugere-se que os alunos utilizem *laptops* em sala de aula, organizada de acordo com a formação e a quantidade de grupos. O caso pode ser adotado em disciplinas de diversos cursos nos quais o docente perceba que sua aplicação contribua com a aprendizagem de temas relacionados a Tecnologias Cívicas, *Data Science* e *Machine Learning*. Para maior detalhamento de recomendação de aplicação, segue a Tabela 1.

Tabela 1 - Disciplinas recomendadas para aplicação do caso para ensino

Nível	Curso	Disciplina Similares
Graduação e Pós-Graduação	Administração, Administração Pública, Contabilidade e afins.	1. Sistemas Gerenciais de Apoio à Decisão 2. Tecnologias da Informação e da Comunicação Aplicadas Aos Negócios
	Análise e Desenvolvimento de Sistemas	1. Gestão da Tecnologia da Informação 2. Probabilidade e Estatística 3. Gestão de Projetos e de Dados
	Ciência da Computação	1. Ciência de Dados e Aprendizagem de Máquina 2. *Big Data* 3. *Machine Learning*
	Engenharia da Computação	1.Gestão de Projetos da Tecnologia da Informação 2.Banco de Dados Probabilidade e Estatística
	Sistemas de Informação	1. Gestão da Tecnologia da Informação 2. Mineração de Dados 3. Inteligência Artificial 4. *Big Data*
Tecnológico	Ciência de Dados	1. Programação Estatística 2. Inteligência Artificial 3. *Machine Learning*
	Bancos de Dados	1. *Big Data* 2. Gestão da Tecnologia da Informação 3. Aplicações para Internet
	Administração	1. Tecnologias da Informação 2. Projeto Integrado de Gestão de Projetos e de Dados 3. Sistemas Integrados de Gestão

Fonte: Elaboração própria (2020)

Na primeira coluna do Quadro 1, os níveis de aplicação estão divididos conforme critérios relacionados às disciplinas destes cursos e o tempo de aplicação do caso. Nesse sentido, para cursos de Graduação e Pós-Graduação tanto *lato sensu* quanto *stricto sensu*, recomenda-se uma aplicação de 2 horas e 50 minutos, divididas seguindo a recomendação abaixo:

1) **Primeiro momento (40 Minutos):**
   Discussão em pequenos grupos e confecção de relatório grupal e procedimentos de decisão. Nesse momento, os relatórios individuais devem ser discutidos e condensados em um só.
2) **Segundo momento (50 minutos):**
   Discussão no grande grupo para exposição de posicionamento e apresentação de novos dilemas surgidos durante a discussão em pequenos grupos. Aqui, aconselha-se que o docente estimule os alunos a buscarem novos dilemas para além daqueles apresentados no enredo do caso.
3) **Terceiro momento (1 hora e 20 minutos):**
   Apresentação dos relatórios de cada grupo e arguição por parte do professor e dos grupos espectadores quanto a solução dos dilemas apresentados no enredo.

Avaliação

Os relatórios devem ser entregues ao docente ao final da aula, e em seu total, somar peso **2 (dois)** na composição da nota final, considerando:

a) Pertinência do conteúdo do relatório
b) Relevância da solução
c) Originalidade dos novos dilemas apresentados

Para os cursos Tecnológicos, dada sua carga horária reduzida, recomenda-se uma abordagem mais breve, de duas horas de duração. Contudo, os procedimentos devem ser os mesmos que aqueles recomendados para os cursos de Graduação e Pós-graduação, flexibilizando ao professor da disciplina

a divisão e o gerenciamento do tempo para cada etapa da discussão e solução do caso. Os relatórios grupais ainda devem ser entregues ao professor ao final da aula e os alunos devem compartilhar suas análises e decisões em plataforma conjunta para acesso de todos. Recomenda-se o uso do Drive (drive.google.com) ou Dropbox (dropbox.com).

Questões para discussão (Resgatando o Dilema)

O dilema deste caso está no desafio de buscar por relações existentes entre os gastos de deputados federais com verba indenizatória e diversas outras variáveis que, segundo o histórico de práticas e denúncias, levam à desconfiança de atos de corrupção, como:

a) Levantamento do CNPJ das empresas onde os gastos foram feitos. Considerando endereços e localizações e seu cruzamento com a agenda oficial do político.
b) As votações do partido e do deputado.
c) Os nomes dos sócios das empresas onde a verba indenizatória é gasta.
d) Data de abertura e baixa da empresa (se houver).
e) Dados que desviam dos padrões.
f) A frequência de aparecimento de nomes de familiares dos deputados nas empresas onde as verbas são gastas.
g) Possíveis contribuições das empresas nos financiamentos de campanha dos partidos ou do próprio deputado.

Para estimular a busca pelas respostas aos itens levantados no dilema, o docente deve resgatar o relatório apresentado pelos grupos durante a discussão em grande grupo.

Avaliação

Uma planilha deverá ser confeccionada com uma amostra dos dados solicitados nos quesitos de (***a***) à (***g***) e apresentada em sala de aula, com peso atribuído de valor **quatro.**

Identificação da demanda e compreensão do dilema

Esta fase deve considerar as questões sugeridas para discussão pela equipe de Samuel e você como parte dela deverá buscar sua solução.

Formular uma pergunta é uma maneira útil de orientar o processo exploratório de análise de dados. Em particular, uma pergunta ou hipótese bem definidos podem servir como uma ferramenta de redução de dimensão que eliminará variáveis irrelevantes (PENG; MATSUI, 2015).

É importante que o docente esteja em constante comunicação com os grupos, estimulando-os a refletir sobre o dilema e suas exigências. Aqui, o docente pode sugerir que os grupos adotem a técnica do 5W2H:

- **Por quê?** (*Why?*): Por que é importante essa análise para o dilema?
- **Quem?** (*Who?*): Quem iremos analisar?
- **O quê?** (*What?*): O que iremos analisar?
- **Onde?** (*Where?*): A análise estará voltada para que contexto?
- **Quando?** (*When?*): Qual período será considerado para as análises?
- **Como?** (*How?*): Como as análises serão realizadas?
- **Quanto?** (*How much?*): Quais os custos envolvidos na análise?

Procedimentos de coleta de dados

Várias fontes de dados fornecem acesso aberto, de forma localizável, acessível, interoperável e reutilizável, que podem ser apropriados para atividades de recuperação, manipulação, processamento, análise e visualização de conjuntos de dados (DINOV, 2018).

Dado que o problema está compreendido e definido, a etapa seguir está em coletar os dados. Nessa fase, é fundamental entender os tipos de dados a serem trabalhados, coletados com o uso de tecnologias como linguagem SQL, *Web Crawlers* e as *Application Programming Interface* (API´s) dos sites:

- ***Dados internos*** (presentes em bancos de dados, planilhas etc.) **x *Dados Externos*** (bases de dados

públicas ou pagas etc.)
- ***Dados estruturados*** (tabelas dos nossos *datasets*) ***x Dados não-estruturados*** (conteúdos de redes sociais, de sites externos etc.).

Processamento e exploração de dados/Análise de dados

É preciso desenvolver mecanismos para estruturar análise para que esta possa ser feita sistematicamente. Assim, é importante entender a ciência de dados, porque a análise de dados nesta fase é crítica para a estratégia. A análise de dados tem impulsionado negócios e instituições, por isso compreender os conceitos fundamentais é determinante para organizar o pensamento analítico e visualizar oportunidades (CAO, 2017).

Após a coleta dos dados, é preciso executar procedimentos de tratamento. Aqui, a atenção se volta para registros duplicados, *missing values*, formatação não-convencional ou inadequada para a análise, preenchimentos inválidos, inconsistências de cadastros. Esses são alguns dos exemplos de problemas achados nesta fase (BLUM; HOPKROFT; KANNAN, 2020).

Nesse momento, um conhecimento básico de alguma linguagem de programação que auxilie na Análise de Dados é fundamental. Sugerimos a linguagem Python, pela sua eficiência, sintaxe intuitiva e de menor curva de aprendizagem, fácil de depurar, de compreender e de expandir (VANDERPLAS, 2016), além de, ultimamente, ter ganho muito espaço no cenário científico.

No entanto, para cursos mais direcionados ao campo organizacional, sem maior aprofundamento em linguagens de programação, como Administração de Empresas, Administração Pública, Contabilidade e cursos afins, a sugestão se dá para o uso de uma planilha de Excel. Nesta aplicação, o aluno pode na aba **Arquivo**, na guia **Opções**, instalar o suplemento **Análise de Dados**, o qual em seguida se tornará ativo no menu **Dados**. Ativada essa opção, o aluno poderá realizar uma AED confiável, estabelecer correlações, analisar covariâncias, construir modelos de regressão e testar diversas hipóteses quanto aos dados coletados do portal da Câmara dos Deputados.

Para alunos que possuem um conhecimento maior da linguagem Python, nesta fase, a sugestão repousa em três indicações para o docente trabalhar em sala ou no laboratório:
Sugestão 1: A biblioteca Python missingno te ajuda a encontrar valores faltantes/nulos nos seus dados.
Sugestão 2: A biblioteca Python Pandas te ajuda a explorar seus *dataframes* com informações estatísticas, descritivas, histogramas etc.
Sugestão 3: A biblioteca Python Scipy que possui abordagem científica e facilita a manipulação de *arrays*.

Nesta fase do projeto de *Data Science*, a resolução do dilema começa a ganhar corpo, pois as habilidades analíticas do tomador de decisão se intensificam e dele é exigida uma forma criativa de pensar em ideias e hipóteses a serem validadas. A ele também cabe a identificação de padrões interessantes nos dados e a interpretação daquilo que esses padrões significam.

Após resgatar para a turma a importância do dilema do caso, o docente deve direcionar as explicações para o próximo passo, que será a seleção das *features* do cenário, sua implementação e aplicação por meio de modelos estatísticos e de *Machine Learning* para validar hipóteses (Caso necessário, o professor deve abrir um espaço para discutir a formulação de hipóteses).

Bibliotecas como a *scikit-learn*, (http://scikit-learn.org/stable/), do Python, que encapsula vários modelos de Classificação, Regressão, Clusterização, Redução de dimensionalidade etc., são muito utilizadas e trazem respostas válidas e acuradas para o dilema. Na mesma perspectiva, ferramentas como o Excel, para os cursos de Administração, possuem *Add-ins* como o XLSTAT365 e o Data Mining que podem realizar operações de Clusterização, Classificação e Predição.

Comunicação de resultados

Há diversos métodos de visualização de dados, tanto para fase exploratória quanto para a analítica, que são: demonstrações de histogramas, gráficos de correlação e de densidade, gráficos de setores, em barra, linhas e gráficos de dispersão, estratégias para exibir árvores de decisão, dendogramas e gráficos mais complexos de superfície 3D. (Williamson, 2016; Mirkin, 2019). Sugerimos que o professor adote os seguintes critérios como estratégia de visualização:

- **Tipo de dados**: estrutura
- **Tipo de tarefa**: forma de visualização
- **Escalabilidade**: limitações dos dados.
- **Dimensionalidade**: atributos usados
- **Posicionamento:** distribuição de atributos e sua interpretação.
- **Necessidade de investigação:** a questão científica

- A composição dos dados
- A distribuição dos dados
- Contraste ou comparação dos elementos de dados, relações, associação
- Mineração de dados.

A comunicação dos resultados de um projeto de *Data Science* como a Operação Serenata do Amor é o alicerce para a tomada de decisão de quem não participou da fase analítica, mas é peça-chave no processo. Nesta fase, o cientista de dados deve atentar para o contexto e para os dados, como no caso da Serenata, no qual os gastos dos deputados são diários e os dados precisam de uma maior frequência de atualização.

O docente deve estimular os alunos por meio dos seguintes questionamentos:

1) Quais técnicas de visualização exploratória estão disponíveis para interrogar graficamente meus dados específicos?
2) Como examinamos associações e correlações emparelhadas de uma maneira multivariada no conjunto de dados?

Ferramentas de visualização, além da biblioteca Matplotlib (https://matplotlib.org/) do Python, ganham cada vez mais adeptos pela facilidade de uso e de aprendizagem. Algumas delas são PowerBI, Qlik Sense Desktop e Tableau Desktop, que são aplicações proprietárias e têm evoluído de apenas ferramentas de *Business Intelligence* para poderosas plataformas de *Analytics.* Sem deixar de mencionar o Excel, há o Add-in D3.Js Charts que possui uma poderosa e intuitiva forma de manipular dados e criar visualizações criativas e confiáveis.

Avaliação

A composição final da nota se dá pela apresentação, em dashboard, das descobertas dos alunos. Considerando:

a) Relação Gráfico x Texto (Clareza e coerência);
b) Denominação dos gráficos;
c) Sintetização, elegância, escala e estética da visualização;

Links úteis

https://www.camara.leg.br/cota-parlamentar/documentos/publ/2880/2015/5660757.pdf
https://github.com/okfn-brasil/serenata-de-amor
https://dadosabertos.camara.leg.br/
https://serenata.ai/
https://jarbas.serenata.ai/dashboard/chamber_of_deputies/reimbursement/
https://twitter.com/RosieDaSerenata
http://www.dados.gov.br/
http://www.tse.jus.br/transparencia
http://www.dados.gov.br/aplicativo/operacao-serenata-de-amor
https://pandas.pydata.org/
https://matplotlib.org/
https://www.scipy.org/
https://powerbi.microsoft.com/pt-br/desktop/
https://www.qlik.com/pt-br/products/qlik-sense/desktop
https://www.tableau.com/pt-br/products/desktop
https://jupyter.org/

Bibliografia Recomendada

Blum, A., Hopcroft, J., & Kannan, R. (2020). **Foundations of data science**. Cambridge University Press.

Cao, L. (2017). Data science: a comprehensive overview. **ACM Computing Surveys** *(CSUR)*, *50*(3), 1-42.

Cota para o exercício da atividade parlamentar. **Câmara dos Deputados**, 2020. Disponível em:https://www2.camara.leg.br/transparencia/acesso-a-informacao/copy_of_perguntas-frequentes/cota-para-o-exercicio-da-atividade-parlamentar/ Acesso em: 20 de Mar. de 2020.

Dinov, I. D. (2018)**. Data science and predictive analytics: Biomedical and health applications using R.** Springer.

Mirkin, B. (2019). **Core Data Analysis: Summarization, Correlation, and Visualization**. Cham: Springer International Publishing.

Operação Serenata de Amor (2018). In Wikipedia. Acessado em 27 Abr 2020. De: https://pt.wikipedia.org/wiki/Opera%C3%A7%C3%A3o_Serenata_de_Amor

Peng, R. D., & Matsui, E. (2015). The Art of Data Science. **A Guide for Anyone Who Works with Data. Skybrude Consulting**, *LLC*.

Tribunal De Contas Da União (2015). **Secretaria de Fiscalização de Tecnologia da Informação. 5 motivos para a abertura de dados na Administração Pública.** Brasília, 2015. Disponível em: http://portal3.tcu.gov.br/portal/pls/portal/docs/2689107.PDF

Acesso em: 25 abr. 2020.

VanderPlas, J. (2016). **Python data science handbook: Essential tools for working with data**. O'Reilly Media, Inc.

Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and 'real-time'policy instruments. **Journal of Education Policy**, *31*(2), 123-141.