UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
Bacharelado em Sistemas de Informação
Inael Wilson Gonçalves de Laia

APLICAÇÃO DA LINGUAGEM JULIA EM CIÊNCIA DE DADOS

Diamantina, MG
2021

Inael Wilson Gonçalves de Laia

APLICAÇÃO DA LINGUAGEM JULIA EM CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado à Universidade Federal dos Vales do Jequitinhonha e Mucuri - UFVJM - como requisito parcial para a obtenção do grau Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Alessandro Vivas Andrade

Diamantina, MG
2021

Inael Wilson Gonçalves de Laia
Aplicação da Linguagem Julia em Ciência de Dados/ Inael Wilson Gonçalves de Laia. – Diamantina, MG, 2021-
50 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Alessandro Vivas Andrade

Trabalho de Conclusão de Curso - Universidade Federal dos Vales do Jequitinhonha e Mucuri, 2021.

1. Julia. 2. Ciência de Dados. 3. Informação. 4. PISA. 5. Algoritmo.

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

FOLHA DE APROVAÇÃO

Inael Wilson Gonçalves de Laia

APLICAÇÃO DA LINGUAGEM JULIA EM CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como requisitos parcial para conclusão do curso.

Orientador: Alessandro Vivas Andrade

Data de aprovação: 11/05/2021

Prof. Dr. Alessandro Vivas Andrade
Faculdade de Ciências Exatas - UFVJM

Profª. Dra.Claudia Beatriz Berti
Faculdade de Ciências Exatas - UFVJM

Prof. MSc. Rafael Santin
Faculdade de Ciências Exatas - UFVJM

Documento assinado eletronicamente por **Alessandro Vivas Andrade**, **Servidor**, em 13/05/2021, às 16:08, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Santin**, **Servidor**, em 13/05/2021, às 16:46,

conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Claudia Beatriz Berti**, **Servidor**, em 13/05/2021, às 20:23, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufvjm.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0353577** e o código CRC **12B97B3E**.

Referência: Processo nº 23086.004440/2021-71 SEI nº 0353577

Dedico este trabalho à minha família e meus amigos.

AGRADECIMENTOS

Aos meus pais, Janaína e Marcelo, por me presentear com todo o apoio necessário para que eu pudesse completar esse objetivo. A minha irmã Marcella, por estar presente e demonstrar o seu apoio. Aos meus amigos, pela amizade que compartilhamos e pelas dificuldades que enfrentamos juntos. Ao meu orientador Professor Alessandro Vivas, pela paciência, confiança e por ter me dado essa oportunidade. Aos meus professores do curso de Sistemas de Informação da UFVJM, que me ensinaram e me guiaram durante este período de tempo. E a todos que, de forma direta ou indireta contribuíram para que eu pudesse me formar.

RESUMO

O processo de exploração e análise de uma base de dados têm se tornado uma atividade cada vez mais interessante para indivíduos ou empresas que buscam aproveitar esses conjuntos de dados para extração de informações úteis ou de fácil interpretação. Por isso, este trabalho teve como objetivo realizar um estudo sobre Ciência de Dados, área a que se vinculam esses conceitos. Com o auxílio da linguagem de programação Julia, voltada para essa área, aplicaram-se diferentes técnicas de filtragem, exploração e classificação de dados em um subset de dados provenientes do PISA (Programme for International Student Assessment). Este teste visa aferir o desempenho de alunos e é aplicado pela OECD (Organisation for Economic Co-operation and Development).

Palavras-chave: Julia. Ciência de Dados. Informação. PISA. Algoritmo.

ABSTRACT

The process of exploring and analyzing a database has become an increasingly interesting activity for individuals or companies looking to take advantage of these data sets to extract useful or easy to interpret information. Therefore, this work aims to carry out a study on Data Science, an area to which these concepts are linked. With Julia programming language, which focuses on this area, different techniques for filtering, exploring and classification of data were applied to a subset of data from PISA (International Student Assessment Program).This test aims to measure student performance and is applied by OECD (Organization for Economic Cooperation and Development).

Keywords: Julia. Data Science. Information. PISA. Algorithm.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Com a popularização da Internet, a quantidade de dados disponíveis cresce a cada dia, uma vez que, toda ação realizada por um usuário pode gerar uma certa quantidade de dados. Devido a isso, praticamente tudo feito nesse meio tem o potencial de se tornar uma informação. Com essa disponibilidade de dados e possibilidade de acesso, pode ser que alguém tenha o interesse na utilização desse dado como base para a geração de algum outro conhecimento novo.

Muitos dos dados produzidos não são devidamente aproveitados e as informações que poderiam ser extraídas de um conjunto de dados acabam se perdendo no grande volume produzido. Isso pode ocorrer porque, muitas vezes, não são analisados da melhor maneira possível e acabam agregando muito pouco na busca por novas informações ou em uma tomada de decisão.

A partir do incremento da quantidade de dados disponível na rede mundial de computadores, uma nova área do conhecimento tem se tornado cada vez mais importante no processo de extração e análise dos mesmos. Assim a Ciência de Dados visa transformar esse volume bruto em informação fácil e útil às decisões. Segundo Passos (2016), a área de Ciência de Dados pode ser definida como um conjunto de técnicas utilizadas no processamento e análise de dados, com intuito de fornecer informações para decisões inteligentes.

Unindo conceitos e técnicas de outras áreas do conhecimento como estatística, matemática e computação, essa ciência busca oferecer ferramentas que possam ser utilizadas para dar sentido a inúmeros dados disponíveis digitalmente e agregar valor nas pesquisas que envolvem aquela informação.

Assim, esse trabalho, pretende demonstrar e enfatizar a importância da Ciência de Dados, explicar os conceitos que estão atrelados e aplicar suas técnicas em conjuntos de dados selecionados previamente, por meio da linguagem de programação Julia que é voltada para essa área.

2 CIÊNCIA DE DADOS

A área de Ciência de Dados tem ganhado mais notoriedade nos tempos recentes, porém coletar dados para serem analisados não é um conceito novo. Uma das primeiras funções oferecidas por sistemas de informação utilizados em organizações, era o de produção de relatórios para que alguma pessoa tivesse a oportunidade de analisá-los e, a partir disso, identificar itens relevantes para tomada de uma decisão.

Ciência de Dados pode ser definida como atividade interdisciplinar, que concilia principalmente duas grandes áreas: Ciência da Computação e Estatística (OLIVEIRA; GUERRA; MCDONNELL, 2018). Seu conceito ganhou cada vez mais força conforme as tecnologias da informação foram se desenvolvendo e se popularizando, pois, as mesmas aumentaram muito a quantidade de dados produzidos dentro e fora das organizações. Hoje em dia a produção de dados vai muito além de simples relatórios. Com a chegada da Internet, uma interação em uma rede social, uma pesquisa ou uma compra podem gerar dados que podem ser aproveitados. Porém, devido a esse aumento, as técnicas antigas de análise foram perdendo a qualidade, pois eram feitas por pessoas e as mesmas já não conseguiam ser capazes de analisar todo o material até então produzido.

Conjuntos de dados tão grandes como os que podem ser encontrados hoje em dia são denominados Big Data. Segundo Junior (2012), Big data refere-se ao conjunto de dados (dataset) cujo tamanho está além da habilidade de ferramentas típicas de banco de dados em capturar, gerenciar e analisar. Segundo Amaral (2016), a definição se dá também por um conjunto de três a cinco "Vs": volume (tamanho e quantidade de dados), velocidade (dinâmica de crescimento e processamento dos dados), variedade (diversidade de origens, formas e formatos dos dados), veracidade (autenticidade, confiabilidade dos dados) e valor (significados atribuídos aos dados, valor agregado).

O papel de quem trabalha com Ciência de Dados se materializa quando ele obtém acesso a esse conjunto de dados, consegue organizá-los, e, por meio das técnicas e ferramentas disponíveis e adequadas, consegue extrair informações que interessem à organização a qual os dados pertencem. Por isso é importante que ao utilizar essas técnicas, a pessoa que irá fazer a análise entenda a diferença de dados para informações e conhecimento, uma vez que para o uso das ferramentas é importante saber as relações e interações que se quer buscar e os resultados que deseja alcançar. Isso se deve ao fato de, se elas não forem aplicadas de forma correta os resultados não trazem conhecimento algum.

Dados podem ser definidos como fatos coletados e normalmente armazenados. Informação é um dado analisado e com algum significado. O conhecimento é a informação interpretada, entendida e aplicada para um fim (AMARAL, 2016). Apesar de uma definição clara, não é possível apontar exatamente o que é um dado, pois para um indivíduo pode ser informação e para outro não. Sendo assim o que dirá se um dado é informação ou não depende da situação ou do local que ele será aproveitado.

Além do conhecimento sobre dados o Cientista de Dados necessita de uma série de competências interdisciplinares. Para atuar em Ciência de Dados, três domínios de conhecimento se inter-relacionam: i) Programação de Computadores; ii) Estatística e Matemática; e iii) Domínio do Conhecimento (RAUTENBERG; CARMO, 2019).

Conforme definição de Amaral (2016) Ciência de Dados são os processos, modelos e tecnologias que estudam o dado durante todo o seu ciclo de vida. E esse ciclo é executado em cinco etapas: produção (criação do dado em formato digital); armazenamento (guardar o dado em uma base de dados); transformação (transformar o dado em modelo ideal para ser utilizado); análise (execução de procedimentos de extração de informação); e descarte (descartar o dado quando não há mais utilidade).

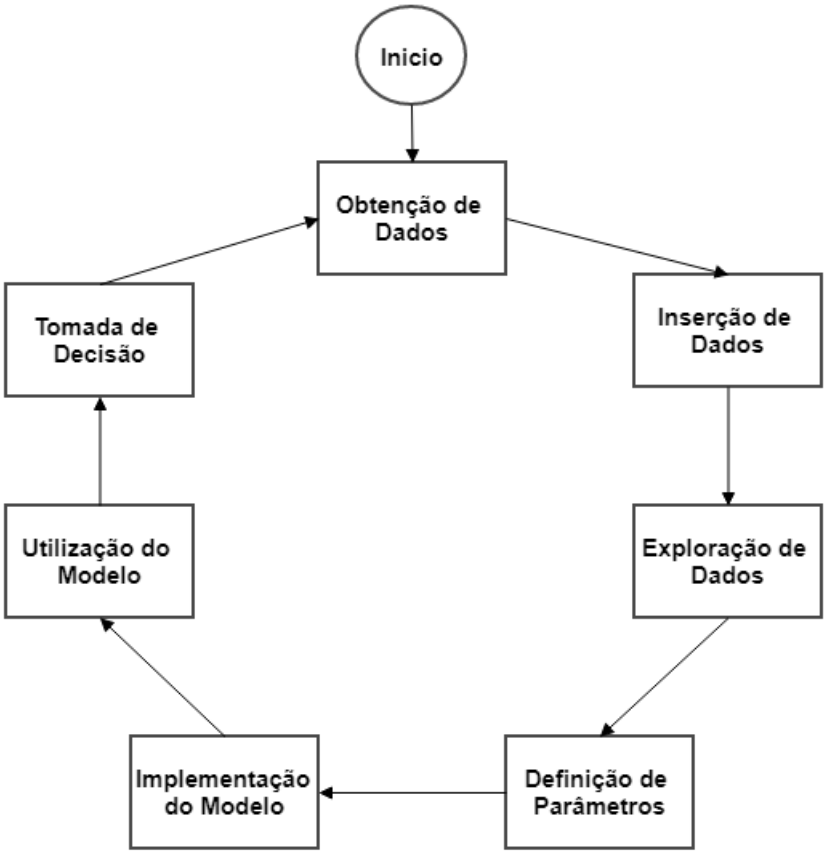
Um conceito atrelado à Ciência de Dados, que pode ser confundido com o mesmo, é o de Mineração de Dados. A princípio ambos podem parecer se referir a mesma coisa. No entanto, é importante enfatizar que o segundo é apenas uma das disciplinas que compõem o primeiro:

> Mineração de Dados é o processo de extração ou mineração de conhecimento em grandes quantidades de dados, e esse é apenas uma parte das diferentes ferramentas à disposição da Ciência de Dados que ainda possui processos vindos de outros campos da computação e da estatística (CÔRTES; PORCARO; LIFSCHITZ, 2002).

Por sua vez Bugnion, Manivannan e Nicolas (2017) apregoam que Ciência de Dados pode ser dividia em um ciclo composto por sete passos (Figura 1):

- **Obtenção de Dados:** Etapa de avaliação e seleção dos dados que serão utilizados durante o processo.
- **Inserção de Dados:** Processo de transformação, organização e centralização dos dados que podem estar vindo de diferentes fontes e dos mais variados formatos.
- **Exploração de Dados:** É feito um estudo preliminar sobre a base de dados disponível para ser possível fazer relações entre os dados e buscar informações.
- **Definição dos Parâmetros:** É implementado um algoritmo de aprendizagem de máquina para tratar os dados e definido os seus parâmetros de entrada e as condições de parada.
- **Implementação do Modelo:** Serão feitos testes e estratégias de treinamento para aperfeiçoar os parâmetros estabelecidos e encontrar o modelo que melhor se adeque a situação.
- **Utilização do Modelo:** Utilização do modelo desenvolvido para que seja encontrado informação no conjunto de dados para poder comprovar sua capacidade.
- **Tomada de Decisão:** A partir dos resultados encontrados no modelo os encarregados tomam uma decisão do que fazer com a informação que foi descoberta.

Figura 1 – Ciclo Ciência de Dados



Fonte: (BUGNION; MANIVANNAN; NICOLAS, 2017). Adaptado.

3 LINGUAGEM DE PROGRAMAÇÃO

Para a realização da análise de dados, faz se necessária a escolha de uma linguagem que possa atender às necessidades da pessoa que executará o processo de manipulação. Existem várias linguagens de programação adequadas a esta análise. Dentre elas, são apresentadas abaixo aquelas que mais se destacam para esta finalidade.

3.1 Tipos de Linguagem

3.1.1 Java

Uma das linguagens de programação mais populares atualmente, Java é uma linguagem orientada ao objeto que pode ser utilizada para inúmeros propósitos que vão além de Ciência de Dados. Foi desenvolvida na primeira metade da década de 90 nos laboratórios da Sun Microsystems com o objetivo de ser mais simples e eficiente do que suas predecessoras (INDRUSIAK, 1996).

No ambiente de Ciência de Dados Java se destaca por suas ferramentas e frameworks para análise de Big Data como o Hadoop e a aplicação de algoritmos de aprendizado de máquina por meio de suas bibliotecas como DL4J, ADAMS e JavaML. Segundo Mehta (2017), Hadoop suporta o processamento de grandes conjuntos de dados em um ambiente distribuído de computação.

Java pode não ser a primeira linguagem refenciada quando o assunto é Ciência de Dados, mas ela oferece boas ferramentas que podem ser úteis para quem trabalha com um foco maior no processamento dos dados.

3.1.2 Julia

Julia é uma linguagem relativamente nova. Ela baseia-se em programação dinâmica de alto nível e foi projetada para atender os requisitos da computação de alto desempenho numérico e científico, sendo também eficaz para a programação em um propósito geral.

Desenvolvida pelos pesquisadores Stefan Karpinski, Jeff Bezanson, Alan Edelman e Viral Shah em 2009, Julia é uma linguagem de programação compilada de código livre de alto nível, projetada com foco na computação científica e numérica de alto desempenho (PEREIRA; SIQUEIRA, 2016).

Julia foi pensada como uma linguagem para computação científica suficientemente rápida, tal como as linguagens C e Fortran, mas igualmente fácil de aprender como o MATLAB, com o objetivo de facilitar a modelagem computacional (PEREIRA; SIQUEIRA, 2016). Apesar de ser muito jovem vem crescendo rapidamente e já oferece abundância de pacotes e bibliotecas prontos para o uso e voltadas para análises de dados, Ciência de Dados e aprendizagem de máquina.

É uma linguagem de programação que está ganhando popularidade em Ciência de Dados por ser moderna, fácil de aprender e rápida com forte capacidade de processamento numérico. Pode ainda não ser tão utilizada quanto outras linguagens mais populares, mas já pode ser considerada uma grande aposta para o futuro.

3.1.3 MATLAB

O MATLAB é uma aplicação especializada para cálculos científicos e de engenharia. Foi projetado para cálculos com matrizes, mas, ao longo dos anos transformou-se em um sistema computacional flexível, capaz de resolver qualquer problema técnico (CHAPMAN, 2015).

Pertence à empresa MathWorks e foi desenvolvido por Cleve Moler em 1970. O programa trabalha em uma linguagem de mesmo nome que oferece ao usuário funções e recursos que buscam facilitar a programação em um ambiente matemático.

Em Ciência de Dados o MATLAB permite ao usuário um ambiente de computação numérica capaz de manipulação de matrizes, visualização de funções e dados, análises estatísticas e várias bibliotecas específicas que podem servir a diversas situações.

3.1.4 Python

É uma das linguagens de programação mais utilizadas para Ciência de Dados devido a sua sintaxe simples fazendo com que a estruturação do código seja feita com facilidade tornando a programação mais produtiva.

Segundo Borges (2014), a linguagem inclui diversas estruturas de alto nível e uma vasta coleção de módulos prontos para uso além de frameworks de terceiros que podem ser adicionados.

Originalmente desenvolvido para servir como ferramenta para físicos e engenheiros, Python foi criado por Guido van Rossum em 1990 e teve como base outra linguagem da época chamada ABC e hoje em dia é uma linguagem de código aberto utilizada e aceita por várias indústrias de alta tecnologia como Google e Disney.

Para Ciência de Dados Python possui um bom número de ferramentas para análise de dados, aprendizagem de máquina e estatística. Nos últimos anos tem ganhado muito espaço no mercado e até rivalizando com o R, mesmo este ainda possuindo uma maior quantidade de ferramentas e facilidades disponíveis ao usuário.

3.1.5 R

Provavelmente a linguagem mais conhecida atualmente devido ao seu foco em análises estatísticas e visualização gráfica de dados. A partir da linguagem S, criada por John Chambers e outros colaboradores,Ross Ihaka e Robert Gentleman na universidade de Auckland, Nova Zelândia, desenvolveram a linguagem R pois acharam que havia a necessidade de um melhor ambiente de software.

R é uma plataforma de análise estatística com ferramentas gráficas muito avançadas. É uma linguagem de software livre e pode ser obtida gratuitamente através de seu site oficial (MAS, 2018).

Suas maiores vantagens estão na facilidade para a produção de gráficos e a possibilidade do uso de expressões matemáticas, quando necessário, e o seu ambiente de desenvolvimento, que possui várias ferramentas que facilitam a manipulação de dados, cálculos e demonstração dos gráficos.

R é uma linguagem de programação que possui um grande foco em manipulação de dados por isso possui uma popularidade tão grande principalmente entre os estatísticos e mineradores de dados.

3.1.6 Scala

Apesar de não ser tão popular entre os programadores quanto, R ou Python no ambiente de Ciência de Dados, a linguagem Scala pode servir como alternativa para aqueles que gostariam de trabalhar com um foco maior em Big Data. Desenvolvida em 2001 por Martin Odersky e por seu grupo de colaboradores da École Polytechnique Fédérale de Lausanne (EPFL) na Suiça.

Scala caracteriza-se por ser uma linguagem multi-paradigma com propósito de aplicação geral. Teve como base de sua criação o Java por isso além de possuir grande similaridade também é compatível com a mesma, porém com uma sintaxe mais simples do que seu antecessor.

Scala é uma boa linguagem para aqueles que buscam trabalhar com Big Data, pois ela possui ferramentas que tentam aproveitar ao máximo a computação distribuída em cluster de computadores. Outro ponto positivo é a abundância de bibliotecas de aprendizagem de máquina e engenharia que essa linguagem possui. Todavia, para aqueles que não irão trabalhar apenas com Big Data Scala não possui análise e visualização de dados (BOBRIAKOV, 2018).

3.1.7 SQL

O Structured Query Language (SQL) é a linguagem padrão para acesso a banco de dados e muito utilizada no meio dos programadores que trabalham com na execução de comandos em banco de dados relacionais.

Sua primeira implementação foi realizada pela IBM em meados de 1970. Ela é baseada na conceituação feita no trabalho do Dr. E.F. Cood sobre banco de dados relacional. Em 1979 a empresa Relational Software Inc., que mais tarde se tornaria a Oracle Corporation, lançou a primeira implementação comercial da linguagem(PRICE, 2009).

O SQL é utilizado para definir estruturas como as tabelas em um banco de dados, faz manipulação de dados por meio de comandos, selecionar, inserir, excluir e atualizar dados no banco de dados e pode também exercer o controle de dados especificando a autorização aos dados do banco.

Em Ciência de Dados o dado é essencial para que o programador possa começar a exercer seu trabalho. O SQL em si, oferece pouco para análise dos dados, mas, em contrapartida, seu entendimento faz com que o programador possa visualizar melhor a base de dados a sua disposição, possa identificar as estruturas e possíveis problemas que possam comprometer a fase de exploração dos mesmos.

3.2 Linguagem escolhida

Como demonstrado acima, há várias opções de linguagens de programação que oferecem um bom suporte para a aplicação das técnicas de Ciência de Dados, porém, para o desenvolvimento deste trabalho, a linguagem a ser utilizada para exploração dos dados do PISA foi Julia.

Apesar de outras linguagens terem uma presença maior no mercado devido ao tempo em que estão em atividade na área, a experiência pessoal com Julia em projetos menores foi determinante para sua escolha. Sua sintaxe é simples e faz com que a sua utilização seja uma boa opção para o desenvolvimento das próximas etapas deste trabalho.

Além disso, possui acesso a uma ferramenta de apoio denominada Jupyter Notebook, que fornece uma aplicação baseada em web, que é útil para capturar todo o processo de computação: desenvolvimento, documentação e execução do código, bem como comunicar os resultados (JUPYTER, 2015). Essa é uma aplicação de código livre desenvolvida pela organização Project Jupyter.

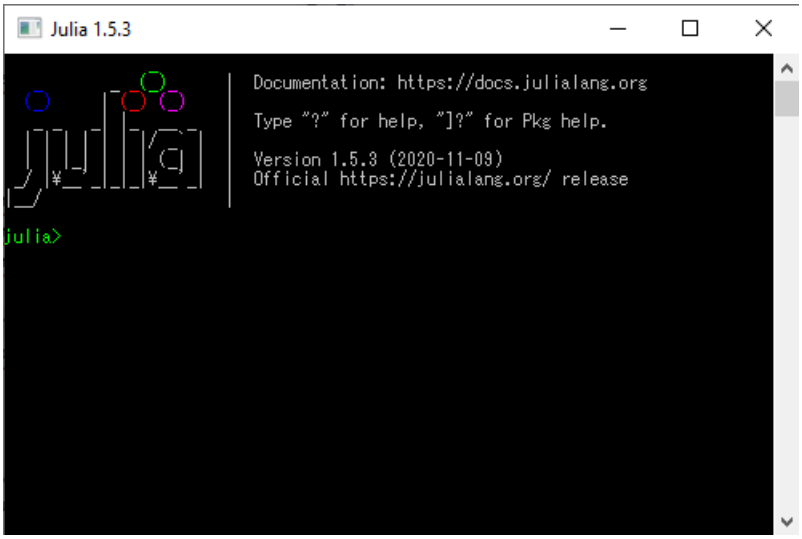
Julia possui um extenso número de bibliotecas com ferramentas que podem auxiliar o trabalho com Ciência de Dados, dentre as quais pode-se destacar as seguintes, que serão utilizados neste trabalho:

- **DataFrames:** É uma biblioteca com ferramentas para se trabalhar com dados tabulares de uso geral e possui integrações com várias outras bibliotecas.
- **CSV:** Biblioteca com ferramentas para o processo de leitura e escrita de base de dados com extensão do tipo (.csv).
- **StatFiles:** Biblioteca com ferramentas para a leitura de conjuntos de dados com extensões (.sav) e (.sas7bdat) gerados por programas estatísticos como SPSS e SAS.
- **Statistics:** Biblioteca com funções básicas de estatística como média, desvio padrão, máximo e mínimo.
- **DataFramesMeta:** Biblioteca com ferramentas que facilitam o processo de manipulação de uma base de dados.
- **Lathe:** É uma biblioteca para modelagem preditiva, com ferramentas para pré-processamento, aprendizado de máquina e validação.

3.2.1 Introdução à linguagem de programação Julia

A interface do Julia é um terminal simples com o logo da linguagem e campo de desenvolvimento em que os códigos devem ser colocados (Figura 2).

Figura 2 – Interface Julia



Fonte: Próprio Autor

Para ter acesso ao ambiente de desenvolvimento Jupyter, faz-se necessário realizar a sua instalação. Para tal, deve-se primeiramente escrever no terminal que será utilizado o gerenciador de pacotes da linguagem denominado *Pkg*. Por meio do comando *add*, o gerenciador possibilita que um usuário instale os pacotes com as ferramentas que ele necessita. Sendo assim, o primeiro pacote que será adicionado é o do *IJulia*, que possui o ambiente de desenvolvimento Jupyter combinado com a linguagem Julia (Figura 3).

Figura 3 – Adicionar IJulia

Fonte: Próprio Autor

Com o pacote IJulia instalado, o programador pode iniciar o Jupyter por meio do comando *notebook* (Figura 4).

Figura 4 – Iniciar Jupyter

Fonte: Próprio Autor

Dentro do ambiente de desenvolvimento o usuário pode criar um novo arquivo para trabalhar com a linguagem (Figura 5).

Figura 5 – Criar Arquivo

Fonte: Próprio Autor

No novo arquivo, o usuário terá o campo no qual ele poderá escrever um código e testá-lo. Nessa nova área, os pacotes podem ser instalados por meio do comando *Pkg.add* e utilizados com o comando *using* da mesma forma que foi realizada anteriormente.

4 APRENDIZADO DE MÁQUINA

Aprendizagem de máquina é o estudo científico de algoritmos e modelos estatísticos que os sistemas de computador usam para realizar uma tarefa específica sem usar instruções explícitas, confiando em padrões e inferência (MORAES, 2019). Pertencente ao campo da inteligência artificial, possui à sua disposição técnicas computacionais capazes de agrupar, classificar dados e resolver problemas de forma automática.

Existem três tipos de aprendizado de máquina para resolução de problemas:

- **Supervisionado:** No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido (MONARD; BARANAUSKAS, 2003).

- **Não Supervisionado:** No aprendizado não supervisionado o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou clusters (MONARD; BARANAUSKAS, 2003).

- **Por Reforço:** No aprendizado por reforço o agente aprende através da exploração do ambiente e das possíveis ações em cada estado após inúmeras iterações, com o objetivo de maximizar a recompensa de suas ações (MORAES, 2019).

Após a escolha de uma linguagem de programação é importante que o cientista de dados tenha conhecimento de alguns dos algoritmos de aprendizado de máquina, uma vez que eles podem ser métodos úteis no processo de análise dos dados.

A seguir serão descritos alguns algoritmos que podem ser utilizados em Ciência de Dados.

4.1 Tipos de Algoritmos

4.1.1 Árvore de decisão

Árvores de Decisão é um tipo de algoritmo supervisionado que é utilizado como método de classificação de dados. Segundo Garcia (2003), esse algoritmo faz a divisão de um conjunto de dados em vários subconjuntos, conhecidos como nós. A classificação ocorre à medida que são percorridos os caminhos por esses nós, até ser encontrado um que contenha a característica determinante do caminho seguido, que é então chamado de folha. A Árvore de Decisão é uma das melhores formas para se identificar a maior parte das relações entre as variáveis que estão sendo manipuladas, além de ser fácil de entender e não necessitar de uma limpeza total do conjunto de dados.

4.1.2 Floresta aleatória

Floresta aleatória é um algoritmo de aprendizagem supervisionada que foca no processo de classificação e regressão de dados e uso do método de Árvore de Decisão. Porém como destaca Neto (2014) no algoritmo de florestas aleatórias, o conjunto de dados é dividido de forma aleatória em subconjuntos menores. Cada um desses conjuntos escolhe um atributo de forma aleatória como nós e a partir disso é formada uma Árvore de Decisão. Com isso pode se dizer que uma Floresta Aleatória nada mais é do que um conjunto de Árvores de Decisões, sendo assim um algoritmo mais poderoso.

4.1.3 KNN (K-vizinhos mais próximos)

O algoritmo de vizinhos mais próximos é um método simples e mais utilizado em classificação, que armazena os casos à sua disposição para classificar novos casos em grupos conforme a maioria de seus vizinhos. Apesar da simplicidade, Costa (2014) destaca que o desempenho desse método está diretamente ligado ao tamanho da base de dados, quanto maior o volume da base, maior será o custo computacional para aplicá-lo.

4.1.4 K-Means

É um algoritmo do tipo não supervisionado e foca na resolução de problemas em clusters. Segundo Guidini (2008), K-Means é um método de agrupamento não-hierárquico por repartição, que consiste num procedimento em que, dado um número de clusters previamente determinado, calculam-se pontos que representam os centros desses clusters. É uma técnica interessante para se fazer a identificação de valores anormais, também chamados de outliers, pois aqueles que não possuem similaridades com outros se destacam na representação feita pelo algoritmo.

4.1.5 Regressão linear

A análise de Regressão Linear pode ser entendida como previsão. O objetivo é prever os valores de uma variável dependente com base em resultados da variável independente(CAVALCANTE; VIANNA, s.d.). Essa técnica pode ser dividida em dois tipos, simples e múltipla, em que a diferença entre esses dois se dá pelo número de variáveis independentes que serão utilizadas apenas uma ou mais de uma.

4.1.6 Regressão logística

A regressão logística é uma metodologia estatística que visa estimar probabilidades de ocorrências em variáveis dependentes do tipo binário (LEITE, 2011). Ou seja, é um método que busca estimar a probabilidade de se um evento pode ou não acontecer.

4.1.7 SVM (Máquina de vetores de suporte)

É uma técnica de aprendizado de máquina que pode ser utilizada para a classificação ou regressão de um conjunto de dados. As SVMs são embasadas pela teoria de aprendizado estatístico (LORENA; CARVALHO, 2007). Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. O algoritmo busca separar o conjunto de dados nas características desejadas e assim facilitar a visualização do mesmo.

5 BASE DE DADOS PISA

Para iniciar a aplicação dos conceitos e técnicas de Ciência de Dados é necessário, junto da escolha de uma linguagem de programação, a seleção de uma base de dados. Neste trabalho a base de dados escolhida foi o PISA (Programme for International Student Assessment) que é uma coletânea de informações sobre o desempenho dos estudantes de 15 anos (INEP, 2018). Esses dados estão disponíveis no site oficial do PISA para o uso de profissionais em estatística e pesquisadores que queiram fazer uma análise do conteúdo do PISA.

Desenvolvido pela OECD (Organisation for Economic Co-operation and Development), o PISA busca testar as habilidades e conhecimentos desses estudantes nas disciplinas de matemática, leitura e ciência. Os resultados permitem que cada país avalie os conhecimentos e as habilidades dos estudantes de seus próprios países em comparação com os de outros países (INEP, 2018).

Outras informações que também são coletadas pelo PISA, e são consideradas importantes são as atitudes e motivações do estudante, além de também avaliar habilidades como resolução colaborativa de problemas. O conteúdo encontrado se baseia no que pode ser encontrado em currículos em volta do mundo e foca na capacidade dos estudantes em aplicar conhecimento, analisar, raciocinar e comunicar efetivamente enquanto examinam, interpretam e resolvem problemas.

O objetivo principal do PISA é informar e dar suporte às decisões educacionais que são feitas em outros países. A pesquisa é realizada a cada três anos, um tempo ideal, segundo eles, para que mudanças e inovações possam demonstrar melhorias ou declínios, bem como as quedas ou melhorias no desempenho dos estudantes possam ser abordadas. Segundo os idealizadores desse teste, a faiza etária de 15 anos é escolhida devido esta ser a idade em que a maior parte dos estudantes dos países selecionados estão próximos do fim de seus estudos. A seleção de alunos para o teste tenta ser a mais inclusiva possível, pois, assim, a amostra de estudantes será mais representativa do todo. Em 2018, 79 países participaram do PISA, sendo 37 deles membros da OECD e 42 países/economias parceiras (INEP, 2018).

A forma principal de aplicação do teste feito pelo PISA se dá na forma de múltipla escolha, haja vista eles acreditarem ser uma forma confiável e eficiente de avaliação. Todavia normalmente, até um terço da avaliação pode ser composta de perguntas abertas. Além das perguntas relacionadas às unidades curriculares, os alunos também devem responder um questionário fornecendo informações sobre si mesmos, suas atitudes em relação à aprendizagem e suas moradias. Os diretores das escolas também participam da pesquisa e respondem um questionário sobre as suas escolas.

Para a criação dos questionáriosa serem aplicados, os desenvolvedores do PISA contam com a participação dos países convidados, os quais enviam suas perguntas que serão adicionadas a itens desenvolvidos por profissionais e contratados da OECD. Essas perguntas são então avaliadas por esses contratados e por outros países participantes. Apenas as perguntas

aceitas de forma unânime, são utilizadas pelo PISA. Após todo esse processo, é feita uma simulação em todos os países participantes e caso uma pergunta seja avaliada como muito fácil ou muito difícil ela é retirada do teste principal.

O PISA conta com a contribuição de convidados e instituições participantes para o seu desenvolvimento, sendo elas:

- Conselho de Administração PISA: Composta por representantes da OECD e membros associados ao PISA. Essa instituição determina as políticas e prioridades e fazem com que elas sejam respeitadas em cada aplicação

- Secretário OECD: Responsável pela manutenção diária do PISA. Monitora a implementação das pesquisas. Gerência de assuntos administrativos do Conselho de Administração PISA. Atua como intermediário entre Conselho de Administração e Consórcio.

- Gerente de Projetos Nacionais do Pisa: São selecionados pelos seus próprios governos e supervisionam a implementação do PISA em cada país participante.

- Consórcio PISA: São profissionais contratados que ficam responsáveis pelo design e implementação das pesquisas. São selecionados pelo Conselho de Administração por meio de concursos internacionais

- Autoridades Educacionais: Ministérios da educação dos países participantes

- Grupos de Especialistas no assunto PISA: Peritos nos três assuntos principais dos testes do PISA, leitura, matemática e ciência. Projetam a estrutura teórica do PISA.

- Grupo de Especialistas do Questionário PISA: Grupo que fornece orientação na construção do questionário.

O financiamento do PISA é feito por meio de contribuições diretas dos países participantes e de autoridades governamentais das economias, geralmente ministérios da educação.

Os resultados do PISA podem ser classificados em escalas específicas desenvolvidas para cada área com o objetivo de mostrar as competências testadas. Essas escalas são divididas em níveis que representam os grupos testados pelo questionário do PISA, começando no nível um com questões que necessitam de habilidades mais básicas para serem completadas e aumentando em dificuldade conforme o nível. Quando corrigida a pontuação do estudante em leitura, matemática e ciências pode ser encontrada na escala apropriada.

O PISA não fornece uma pontuação coletiva para todas as áreas testadas combinadas, ao contrário, fornece uma pontuação para cada assunto e determina as classificações pela pontuação média de cada uma delas. No entanto, não é possível atribuir uma única classificação exata em cada assunto a cada país, uma vez que o PISA testa apenas uma amostra de estudantes de cada país e esse resultado é ajustado para refletir toda a população de estudantes de 15 anos daquele país.

5.1 Questionários

Além das avaliações em ciência, leitura e matemática, o PISA aplica questionários contextuais a estudantes, pais, diretores de escolas e professores. A seguir serão descritos os assuntos abordados e objetivos dos questionários aplicados a cada um dos participantes envolvidos no programa.

5.1.1 Estudantes

Questionários direcionados aos estudantes podem ser questões mais diretas com o objetivo de traçar o perfil do aluno, o de sua família, de sua moradia e da escola. E também podem ser perguntas mais pessoais, como a visão do aluno sobre as atividades realizadas na escola, seu tempo de aprendizado, importância da leitura, motivações e objetivos de vidas.

5.1.2 Pais

Os pais respondem um questionário similar ao aplicado aos estudantes, com perguntas sobre a família, escola e o ambiente em que estão inseridos. O objetivo deste questionário é ter a visão dos pais sobre esses assuntos e tentar entender as influências que o estudante pode ter no seu desenvolvimento.

5.1.3 Diretores

Os diretores respondem questões que providenciam informações sobre a escola, seu gerenciamento, professores e alunos. O objetivo deste questionário é traçar similaridades entre as escolas e entender o contexto em que os estudantes estão inseridos.

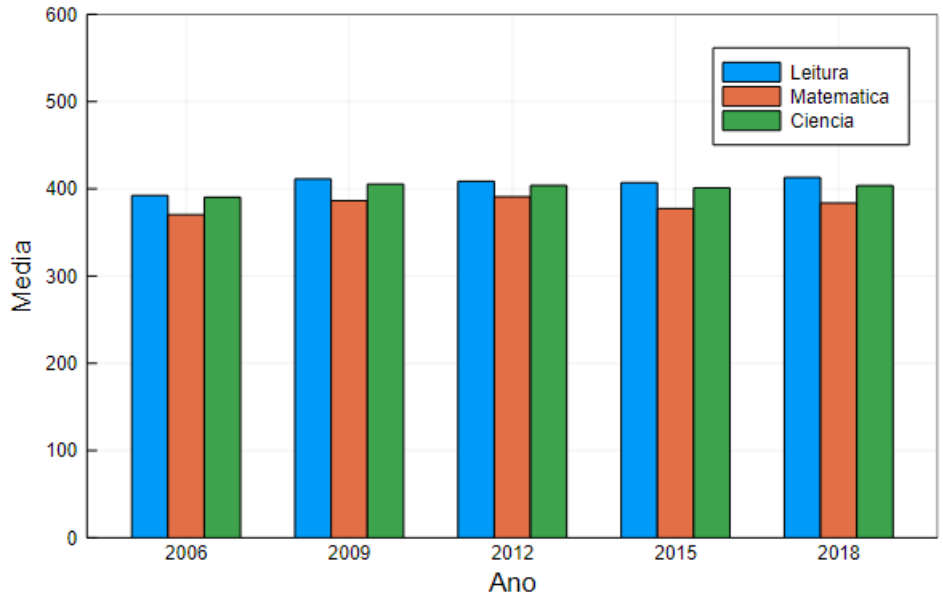
5.1.4 Professores

Os professores responderam a questões que dizem respeito a escola, qualificação e desenvolvimento profissional, práticas de ensino, ambiente para aprendizagem, relação com os pais dos estudantes. O objetivo desse teste é traçar o perfil dos professores e com isso relacioná-lo ao desenvolvimento dos estudantes.

5.2 Pisa no Brasil

O Brasil participa do PISA desde sua primeira edição no ano 2000, sendo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) o órgão responsável pelo planejamento e operacionalização dessa avaliação no país (INEP, 2018). No ano de 2018 os estudantes brasileiros conseguiram uma média de 413 pontos em Leitura, 384 pontos em Matemática e 404 pontos em Ciências. Para efeito de comparação com outros anos, as notas dos estudantes brasileiros podem ser visualizadas a seguir (Figura 6).

Figura 6 – Notas médias do Brasil no PISA



Fonte: Próprio Autor

6 EXPLORAÇÃO DOS DADOS

A partir desse capítulo, o processo de exploração e filtragem dos dados será demonstrado utilizando a linguagem Julia e outras ferramentas descritas anteriormente.

6.1 Importação dos Dados

Para iniciar o processo de exploração, a primeira etapa a ser realizada é a importação das bases de dados. Existem diferentes tipos de arquivos que podem ser importados e a forma com que esse procedimento deve ser feito varia com o tipo de arquivo.

O primeiro conjunto de dados a ser utilizado é um arquivo do tipo *csv* que possui os dados do desempenho médio dos estudantes. Para isso, será necessário a instalação de dois novos pacotes, o *DataFrames* que oferece ferramentas para se trabalhar com dados tabulares, e *CSV*, que oferece ferramentas de leitura e escrita nesse tipo específico de arquivo. Além disso, também será utilizado os comandos *nrow* e *ncol* para demonstrar o tamanho de linhas e colunas do arquivo, *typeoff* que diz o tipo do arquivo, e por fim o comando *describe* que faz a sumarização dos dados (Figura 7).

Figura 7 – Dados-CSV

```
using DataFrames
using CSV
leitura = CSV.read("Reading.csv", DataFrame);
println("Linhas: ", nrow(leitura))
println("Colunas: ", ncol(leitura))
println("Tipo: ",typeof(leitura))
```

```
Linhas: 787
Colunas: 8
Tipo: DataFrame
```

```
describe(leitura)
```

3 rows × 8 columns

	variable	mean	min	median	max	nunique	nmissing	eltype
	Symbol	Union...	Any	Union...	Any	Union...	Union...	DataType
1	LOCATION		AUS		USA	47		String
2	INDICATOR		PISAREAD		PISAREAD	1		String
3	SUBJECT		BOY		TOT	3		String
4	MEASURE		MEANSCORE		MEANSCORE	1		String
5	FREQUENCY		A		A	1		String
6	TIME	2009.62	2000	2009.0	2018			Int64
7	Value	486.163	358.0	493.0	574.0			Float64
8	Flag Codes						787	Missing

Fonte: Próprio Autor

Em seguida, serão realizadas as mesmas etapas, porém com um arquivo do tipo *sav*. Essa base de dados possui as respostas do questionário dos professores, o qual pode ser importado com o auxílio do pacote *StatFiles* que possui ferramentas para leitura de outros tipos de dados (Figura 8).

Figura 8 – Dados-SAV

```
using StatFiles
prof = DataFrame(load("QProfessor.sav"));
println("Linhas: ", nrow(prof))
println("Colunas: ", ncol(prof))
println("Tipo: ",typeof(prof))
```

```
Linhas: 107367
Colunas: 352
Tipo: DataFrame
```

```
describe(prof)
```

	variable	mean	min	median	max	nunique	nmissing	eltype
	Symbol	Union…	Any	Union…	Any	Union…	Int64	Union
1	CNTRYID	482.54	8.0	591.0	840.0		0	Union{Missing, Float64}
2	CNT		ALB		USA	19	0	Union{Missing, String}
3	CNTSCHID	4.82555e7	800002.0	5.91001e7	8.40002e7		0	Union{Missing, Float64}
4	CNTTCHID	4.82587e7	800001.0	5.91007e7	8.40038e7		0	Union{Missing, Float64}
5	TEACHERID	4.69419	4.0	5.0	5.0		16177	Union{Missing, Float64}
6	CYC		07MS		07MS	1	0	Union{Missing, String}
7	NatCen		000800		084000	19	0	Union{Missing, String}
8	Region	48256.8	800.0	59100.0	84000.0		0	Union{Missing, Float64}
9	STRATUM		ALB0101		USA0308	311	0	Union{Missing, String}
10	SUBNATIO		0080000		8400000	19	0	Union{Missing, String}
11	OECD	0.44482	0.0	0.0	1.0		0	Union{Missing, Float64}

Fonte: Próprio Autor

6.2 Dados Selecionados

Devido ao seu tamanho, para se trabalhar com o questionário dos professores um grupo de questões foi selecionado para o processo realizado. Todos os anos em que é realizado, o PISA muda o foco da sua avaliação. Em 2015, por exemplo, o foco foi em ciência, mas, agora em 2018, ele teve como foco a leitura. Sendo assim, a filtragem foi realizada tendo como base os resultados dos alunos em leitura. Já para os professores, foram selecionados as seguintes questões, que também se relacionam com o tema:

(TC155) Com que frequência você ensina os seguintes aspectos da compreensão de leitura em suas aulas?

(TC155Q02HA) Estratégias de Resumo.

(TC155Q03HA) Conectar textos com conhecimento de conteúdo anterior.

(TC164Q01HA) Durante este ano, quantas páginas tinha o maior texto que os alunos tiveram que ler para as aulas?

(TC172Q01HA) Qual das seguintes afirmações melhor descreve como você lê livros (sobre qualquer assunto)?

6.2.1 Desempenho médio dos alunos em leitura

Os dados a serem utilizados correspondem a base de dados com o desempenho dos alunos em leitura. O PISA ainda fornece duas outras bases de dados com os resultados dos países nos temas de matemática e ciência. A base de dados das notas médias em leitura possui, inicialmente, oito colunas e 787 linhas. Cada país possui duas médias para cada escola participante, uma para estudantes do sexo masculino, outra para feminino em cada um dos anos que o PISA foi aplicado (Figura 9).

Figura 9 – Desempenho Médio

```
leitura
```

787 rows × 8 columns

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
	String	String	String	String	String	Int64	Float64	Missing
1	AUS	PISAREAD	BOY	MEANSCORE	A	2000	513.0	*missing*
2	AUS	PISAREAD	BOY	MEANSCORE	A	2003	506.0	*missing*
3	AUS	PISAREAD	BOY	MEANSCORE	A	2006	495.0	*missing*
4	AUS	PISAREAD	BOY	MEANSCORE	A	2009	496.0	*missing*
5	AUS	PISAREAD	BOY	MEANSCORE	A	2012	495.09	*missing*
6	AUS	PISAREAD	BOY	MEANSCORE	A	2015	487.0	*missing*
7	AUS	PISAREAD	BOY	MEANSCORE	A	2018	487.0	*missing*
8	AUS	PISAREAD	GIRL	MEANSCORE	A	2000	546.0	*missing*
9	AUS	PISAREAD	GIRL	MEANSCORE	A	2003	545.0	*missing*
10	AUS	PISAREAD	GIRL	MEANSCORE	A	2006	532.0	*missing*
11	AUS	PISAREAD	GIRL	MEANSCORE	A	2009	533.0	*missing*
12	AUS	PISAREAD	GIRL	MEANSCORE	A	2012	529.542	*missing*
13	AUS	PISAREAD	GIRL	MEANSCORE	A	2015	519.0	*missing*
14	AUS	PISAREAD	GIRL	MEANSCORE	A	2018	519.0	*missing*
15	AUT	PISAREAD	BOY	MEANSCORE	A	2000	476.0	*missing*

Fonte: Próprio Autor

No desenvolvimento deste trabalho, foram utilizadas as médias de 2018, as quais foram obtidas a partir do seguinte comando (Figura 10).

Figura 10 – Ano 2018

```
leitura = leitura[leitura[!,:TIME] .== 2018, :];
```

```
leitura = select!(leitura, :LOCATION, :SUBJECT, :Value)
```

123 rows × 3 columns

	LOCATION	SUBJECT	Value
	String	String	Float64
1	AUS	BOY	487.0
2	AUS	GIRL	519.0
3	AUT	BOY	471.0
4	AUT	GIRL	499.0
5	BEL	BOY	482.0
6	BEL	GIRL	504.0
7	CAN	BOY	506.0
8	CAN	GIRL	535.0
9	CZE	BOY	474.0
10	CZE	GIRL	507.0
11	DNK	BOY	486.0
12	DNK	GIRL	516.0

Fonte: Próprio Autor

Na próxima etapa, para se obter as médias gerais e outros valores estatísticos de cada localidade, foi necessário o agrupamento dos alunos de cada país. Utilizou-se o comando *groupby* que teve como princípio dividir as linhas da base de dados em grupos de acordo com o valor da coluna estabelecida (Figura 11).

Figura 11 – Agrupamento

```
In [19]: g = groupby(leitura, :LOCATION)
```

Out[19]: **GroupedDataFrame with 41 groups based on key: LOCATION**

First Group (2 rows): LOCATION = "AUS"

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
	String	String	String	String	String	Int64	Float64	Missing
1	AUS	PISAREAD	BOY	MEANSCORE	A	2018	487.0	*missing*
2	AUS	PISAREAD	GIRL	MEANSCORE	A	2018	519.0	*missing*

⋮

Last Group (2 rows): LOCATION = "LTU"

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
	String	String	String	String	String	Int64	Float64	Missing
1	LTU	PISAREAD	BOY	MEANSCORE	A	2018	457.0	*missing*
2	LTU	PISAREAD	GIRL	MEANSCORE	A	2018	496.0	*missing*

Fonte: Próprio Autor

Por fim, utilizou-se o comando *combine*, para unificar as linhas de cada grupo formado, e as funções estatísticas, encontradas no pacote *Statistics*, para se obter as estatísticas desejadas. Com as notas médias dos estudantes do sexo masculino e feminino de cada país foi formado uma nova base de dados com os valores da média geral de cada país, o desvio padrão, o valor mínimo e o máximo (Figura 12).

Figura 12 – Estatísticas por País

```
using Statistics
a = combine(g, :Value => mean => :media, :Value => std => :desvio, :Value => minimum => :min, :Value => maximum => :max);
withenv("LINES" => 41) do
    display(a)
end
```

41 rows × 5 columns

	LOCATION	media	desvio	min	max
	String	Float64	Float64	Float64	Float64
1	AUS	503.0	22.6274	487.0	519.0
2	AUT	485.0	19.799	471.0	499.0
3	BEL	493.0	15.5563	482.0	504.0
4	CAN	520.5	20.5061	506.0	535.0
5	CZE	490.5	23.3345	474.0	507.0
6	DNK	501.0	21.2132	486.0	516.0
7	FIN	520.5	36.0624	495.0	546.0
8	FRA	492.5	17.6777	480.0	505.0
9	DEU	499.0	18.3848	486.0	512.0
10	GRC	458.0	29.6985	437.0	479.0
11	HUN	476.0	18.3848	463.0	489.0
12	ISL	474.0	28.2843	454.0	494.0
13	IRL	518.0	16.9706	506.0	530.0

Fonte: Próprio Autor

6.2.2 Questionário professores

A base de dados com as respostas dos professores possui 352 colunas e 107367 linhas. Armazena as respostas individuais dos professores participantes de cada país (Figura 13).

Figura 13 – Questionário Professores

```
In [23]: prof
```

Out[23]: 107,367 rows × 352 columns (omitted printing of 343 columns)

	CNTRYID	CNT	CNTSCHID	CNTTCHID	TEACHERID	CYC	NatCen	Region	STRATUM
	Float64?	String?	Float64?	Float64?	Float64?	String?	String?	Float64?	String?
1	8.0	ALB	800057.0	800001.0	5.0	07MS	000800	800.0	ALB0203
2	8.0	ALB	800121.0	800002.0	5.0	07MS	000800	800.0	ALB0107
3	8.0	ALB	800140.0	800003.0	5.0	07MS	000800	800.0	ALB0101
4	8.0	ALB	800149.0	800004.0	5.0	07MS	000800	800.0	ALB0211
5	8.0	ALB	800095.0	800005.0	5.0	07MS	000800	800.0	ALB0204
6	8.0	ALB	800151.0	800006.0	4.0	07MS	000800	800.0	ALB0105
7	8.0	ALB	800024.0	800007.0	5.0	07MS	000800	800.0	ALB0204
8	8.0	ALB	800231.0	800008.0	5.0	07MS	000800	800.0	ALB0203
9	8.0	ALB	800302.0	800009.0	5.0	07MS	000800	800.0	ALB0101
10	8.0	ALB	800200.0	800010.0	5.0	07MS	000800	800.0	ALB0101
11	8.0	ALB	800143.0	800011.0	5.0	07MS	000800	800.0	ALB0107
12	8.0	ALB	800092.0	800012.0	5.0	07MS	000800	800.0	ALB0203
13	8.0	ALB	800095.0	800013.0	5.0	07MS	000800	800.0	ALB0204
14	8.0	ALB	800075.0	800014.0	5.0	07MS	000800	800.0	ALB0203

Fonte: Próprio Autor

Foram utilizadas apenas questões relacionadas à leitura. Com o comando *select* realizou-se a seleção das colunas a serem utilizadas. Ao final desse procedimento a base de dados possui cinco colunas e 107367 linhas (Figura 14).

Figura 14 – Questões Selecionadas

```
In [5]: prof = select!(prof, :CNT, :TC155Q02HA, :TC155Q03HA, :TC164Q01HA, :TC172Q01HA);
        prof
```

Out[5]: 107,367 rows × 5 columns

	CNT	TC155Q02HA	TC155Q03HA	TC164Q01HA	TC172Q01HA
	String?	Float64?	Float64?	Float64?	Float64?
1	ALB	3.0	3.0	2.0	2.0
2	ALB	3.0	3.0	4.0	2.0
3	ALB	4.0	3.0	2.0	3.0
4	ALB	3.0	4.0	2.0	2.0
5	ALB	4.0	4.0	2.0	2.0
6	ALB	2.0	4.0	2.0	2.0
7	ALB	3.0	4.0	2.0	4.0
8	ALB	3.0	4.0	4.0	2.0
9	ALB	4.0	4.0	2.0	2.0
10	ALB	4.0	4.0	2.0	4.0
11	ALB	2.0	3.0	5.0	2.0
12	ALB	3.0	3.0	2.0	2.0
13	ALB	4.0	4.0	2.0	2.0
14	ALB	3.0	4.0	2.0	4.0
15	ALB	4.0	4.0	6.0	4.0
16	ALB	4.0	4.0	2.0	4.0

Fonte: Próprio Autor

7 CLASSIFICAÇÃO

Para visualização e melhor entendimento dos dados é interessante saber se aquela informação representa algo interessante ou desinteressante. Por isso, é importante, após a seleção dos dados, verificar se os mesmos já estão padronizados de alguma forma que facilite o seu entendimento, e, caso contrário, fazer a classificação.

7.1 Notas médias dos alunos

As notas dos alunos não se encontram classificadas, por isso foi utilizada para essa etapa as escalas de proficiência do próprio PISA. As informações em cada nível são usadas para desenvolver descrições resumidas dos tipos de alfabetização em leitura associada a diferentes níveis de proficiência (PISA, 2018).

- **Level 6 (Nota acima de 698.32):** É exigido do leitor várias inferências, comparações e contrastes detalhados e precisos. Exigem a demonstração de uma compreensão detalhada de um ou mais textos e podem envolver informações de mais de um texto.
- **Level 5 (Maior que 625.61 e menor igual a 698.32):** As tarefas neste nível que envolvem a recuperação de informações requerem que o leitor localize e organize várias informações profundamente incorporadas, inferindo quais partes do texto são relevantes.
- **Level 4 (Maior que 552.89 e menor igual a 625.61):** As tarefas neste nível que envolvem a recuperação de informações requerem que o leitor localize e organize várias informações incorporadas.
- **Level 3 (Maior que 480.18 e menor igual a 552.89):** As tarefas neste nível exigem que o leitor localize e, em alguns casos, reconheça a relação entre várias informações que devem atender a várias condições.
- **Level 2 (Maior que 407.47 e menor igual a 480.18):** Algumas tarefas neste nível exigem que o leitor localize uma ou mais informações, que pode precisar ser inferido ou atender a várias condições.
- **Level 1a (Maior que 334.75 e menor igual a 407.47):** As tarefas neste nível requerem que o leitor compreenda o significado literal das frases ou passagens curtas.
- **Level 1b (Maior que 262.04 e menor igual a 334.75):** As tarefas neste nível requerem que o leitor compreenda o significado literal das frases dentro de uma passagem curta única.
- **Level 1c (189.33 até menor igual a 262.04):** As tarefas neste nível requerem que o leitor compreenda o significado literal de cada palavra escrita e frases em passagens sintaticamente simples e muito curtas com contextos.

Com base na escala do PISA, a classificação foi executada por meio de uma das ferramentas de manipulação do pacote *DataFramesMeta*. O procedimento teve inicio com a seleção das colunas a serem mantidas por meio do comando *select*. A partir disso, foi utilizado o comando *@linq*, que suporta o encadeamento de funções para manipulação de base de dados. Em seguida, com o *transform* foi gerado uma nova coluna denominada *Level* que foi preenchida conforme os valores da coluna *Value* (Figura 15).

Figura 15 – Classificação Alunos

```
using DataFramesMeta
leitura = select!(leitura, :LOCATION, :SUBJECT, :Value);
b = @linq leitura |>
    transform(Level = ifelse.(:Value .> 698.32, "L6", ifelse.((:Value .>  625.61) .& (:Value .<=  698.32), "L5",
            ifelse.((:Value .>  552.89) .& (:Value .<= 625.61), "L4", ifelse.((:Value .>  480.18) .& (:Value .<= 552.89), "L3",
                    ifelse.((:Value .> 407.47) .& (:Value .<= 480.18), "L2", ifelse.((:Value .> 334.75) .& (:Value .<= 407.47), "L1",
                            ifelse.((:Value .> 262.04) .& (:Value .<= 334.75), "L1", ifelse.((:Value .> 189.33) .& (:Value .<= 262.04), "L1", "L1")))))))))
b = filter(row -> !(row.SUBJECT == "TOT" ),  b)
b
```

82 rows × 4 columns

	LOCATION	SUBJECT	Value	Level
	String	String	Float64	String
1	AUS	BOY	487.0	L3
2	AUS	GIRL	519.0	L3
3	AUT	BOY	471.0	L2
4	AUT	GIRL	499.0	L3
5	BEL	BOY	482.0	L3
6	BEL	GIRL	504.0	L3
7	CAN	BOY	506.0	L3
8	CAN	GIRL	535.0	L3
9	CZE	BOY	474.0	L2
10	CZE	GIRL	507.0	L3
11	DNK	BOY	486.0	L3
12	DNK	GIRL	516.0	L3

Fonte: Próprio Autor

7.2 Respostas dos professores

As questões selecionadas para os professores já possuem respostas que seguem uma escala numérica. A quantidade de opções que podem ser assinaladas variam de questão para questão assim como o significado de cada número. Sendo assim, a seguir serão listados o que cada número significa para a questão a qual está atrelado.

As questões TC155 é um conjunto de perguntas sobre assuntos relacionados entre si. Para esse trabalho foram selecionadas as respostas referentes a dois desses assuntos com os respectivos códigos TC155Q02HA e TC155Q03HA. Portanto, as respostas possuem a mesma variação de um a quatro sendo que cada uma dessas respostas possuem o seguinte significado:

- **1:** Nunca ou quase nunca
- **2:** Algumas lições
- **3:** Várias lições
- **4:** Todas ou quase todas as lições

Já a questão de código TC164Q01HA possibilita um número de respostas maiores, nesse caso de um a seis. Como não há relação direta com nenhuma das outras questões, esses números possuem seus próprios significados listados a seguir:

- **1:** Uma página ou menos
- **2:** Entre 2 e 10 páginas
- **3:** Entre 11 e 50 páginas
- **4:** Entre 51 100 páginas
- **5:** Entre 101 500 páginas
- **6:** Mais de 500 páginas

Por fim, a questão de código TC172Q01HA retorna com uma variação de um a quatro para respostas, mas, não possui relação com as primeiras questões. O significado de cada uma das possíveis respostas para essa questão são os seguintes:

- **1:** Eu raramente ou nunca leio livros.
- **2:** Eu leio livros em formato de papel mais regularmente
- **3:** Eu leio livros em formato digital mais regularmente
- **4:** Eu leio livros em formato de papel e digital na mesma quantidade

8 DADOS FALTANTES

Após o processo de importação, seleção e a classificação dos dados, os resultados até o presente momento, apresentaram objetos incompletos que podem interferir em uma eventual extração de informação. Segundo Veroneze (2011), em um conjunto de dados, um objeto (registro ou caso) é completo se todos os seus atributos estão preenchidos com dados apropriados. Um dado faltante indica que um atributo de um objeto está vazio.

Os dados faltantes, que são representados normalmente como valores *missing*, podem aparecer na base de dados por diferentes motivos e é importante saber fazer o tratamento adequado a cada tipo, para que assim a qualidade da amostra não diminua. Os dados faltantes que normalmente aparecem podem ser classificados em uma das três seguintes categorias:

- **Missing Completely at Random (MCAR):** Os dados faltantes não são diferentes dos dados normais. Neste caso, os dados faltantes surgiram de maneira aleatória, e o problema gerado pelos dados faltantes é a perda de poder da análise a ser realizada (ASSUNÇÃO, 2012).

- **Missing at Random (MAR):** Os dados faltantes dependem das variáveis preenchidas, portanto, podem ser totalmente explicadas pelas demais variáveis presentes no banco de dados (ASSUNÇÃO, 2012).

- **Missing Not at Random (MNAR):** Nesta situação os dados faltantes são gerados de forma não mensurável, ou seja, eles dependem de eventos que o pesquisador não consegue observar e controlar (ASSUNÇÃO, 2012).

8.1 Dados faltantes nas bases escolhidas

A seguir, serão demonstrados como identificar e observar a quantidade de dados faltantes em cada coluna de uma base de dados por meio da linguagem Julia.

8.1.1 Alunos

O comando *describe* é uma função que retorna valores estatísticos referentes a uma base de dados. Um desses valores é o número de dados faltantes de cada coluna. Sendo assim, através da realização deste comando, foi possível visualizar a quantidade exata que cada uma das colunas da base de dados das médias dos alunos possuíam.

Figura 16 – Dados faltantes base alunos

```
In [9]: describe(b, :nmissing)
```

Out[9]: 4 rows × 2 columns

	variable	nmissing
	Symbol	Nothing
1	LOCATION	
2	TIME	
3	Value	
4	Level	

Fonte: Próprio Autor

Como pode ser observado na Figura 16, nenhuma das colunas de interesse para esse estudo possui dados faltantes, portanto não será necessário ser feito mais nenhuma manipulação.

8.1.2 Professores

Diferentemente da base dos alunos, utilizando o mesmo comando, dados faltantes puderam ser observados desta vez. Sendo assim, para melhor demonstrar quantos desse tipo de dados estão no conjunto, foi decidido calcular o percentual dos dados faltantes (Figura 17).

Figura 17 – Dados faltantes base professores

```
p = describe(prof, :nmissing)
o=nrow(prof)
o = begin
    p[!, :per] = (p[!,:nmissing] .* 100)/o;
    end
select!(p, :variable, :nmissing, :per)
```

5 rows × 3 columns

	variable	nmissing	per
	Symbol	Int64	Float64
1	CNT	0	0.0
2	TC155Q02HA	18098	16.8562
3	TC155Q03HA	18249	16.9968
4	TC164Q01HA	24520	22.8376
5	TC172Q01HA	17710	16.4948

Fonte: Próprio Autor

Com o objetivo de obter-se uma melhor visualização sobre o número de dados faltantes, decidiu-se agrupar a base de dados por países com o comando *groupby*. Em seguida, foi utilizado o comando *combine* para juntar a base de dados agrupada com os resultados encontrados pela função *describe*. Por fim, com o comando *select* é feita a seleção das colunas que devem ser visualizadas (Figura 18).

Figura 18 – Dados faltantes por país e questões

```
In [139]:
h = groupby(prof, :CNT)
h = combine(describe, h)
select(h, :CNT, :variable, :nmissing)
```

Out[139]: 95 rows × 3 columns

	CNT	variable	nmissing
	String?	Symbol	Int64
1	ALB	CNT	0
2	ALB	TC155Q02HA	166
3	ALB	TC155Q03HA	163
4	ALB	TC164Q01HA	160
5	ALB	TC172Q01HA	152
6	QAZ	CNT	0
7	QAZ	TC155Q02HA	2102
8	QAZ	TC155Q03HA	2080
9	QAZ	TC164Q01HA	2152
10	QAZ	TC172Q01HA	1993
11	BRA	CNT	0
12	BRA	TC155Q02HA	1985
13	BRA	TC155Q03HA	1997
14	BRA	TC164Q01HA	2194
15	BRA	TC172Q01HA	1965

Fonte: Próprio Autor

Para o tratamento desses dados, foi considerado que o objeto se encontra com dados faltantes, pois o professor não respondeu à questão. Sendo assim, todos os valores não respondidos foram representados pelo número zero e essa etapa foi realizada conforme os comandos utilizados abaixo (Figura 19):

Figura 19 – Substituição dos dados faltantes

```
In [13]:
prof[:TC155Q02HA] = collect(Missings.replace(prof[:TC155Q02HA], 0));
prof[:TC155Q03HA] = collect(Missings.replace(prof[:TC155Q03HA], 0));
prof[:TC164Q01HA] = collect(Missings.replace(prof[:TC164Q01HA], 0));
prof[:TC172Q01HA] = collect(Missings.replace(prof[:TC172Q01HA], 0));
prof
```

Out[13]: 107,367 rows × 5 columns

	CNT	TC155Q02HA	TC155Q03HA	TC164Q01HA	TC172Q01HA
	String?	Float64	Float64	Float64	Float64
1	ALB	3.0	3.0	2.0	2.0
2	ALB	3.0	3.0	4.0	2.0
3	ALB	4.0	3.0	2.0	3.0
4	ALB	3.0	4.0	2.0	2.0
5	ALB	4.0	4.0	2.0	2.0
6	ALB	2.0	4.0	2.0	2.0
7	ALB	3.0	4.0	2.0	4.0
8	ALB	3.0	4.0	4.0	2.0
9	ALB	4.0	4.0	2.0	2.0

Fonte: Próprio Autor

Além da substituição, eliminou-se por completo algumas linhas. Isso se deve ao fato de que um ou mais professores de algum lugar pode não ter respondido nenhuma das questões selecionadas, assim, sua inclusão é completamente irrelevante. Com o comando *filter* pode-se realizar a cópia dos dados por meio de uma condição: a que linha apresentasse o valor zero em todas as suas colunas, não seria copiada para a nova base de dados (Figura 20).

Figura 20 – Retirar dados sem respostas

```
q = filter(row -> !(row.TC155Q02HA == 0 && row.TC155Q03HA == 0 && row.TC164Q01HA == 0 && row.TC172Q01HA == 0),  prof)
```

89,968 rows × 5 columns

	CNT	TC155Q02HA	TC155Q03HA	TC164Q01HA	TC172Q01HA
	String?	Float64	Float64	Float64	Float64
1	ALB	3.0	3.0	2.0	2.0
2	ALB	3.0	3.0	4.0	2.0
3	ALB	4.0	3.0	2.0	3.0
4	ALB	3.0	4.0	2.0	2.0
5	ALB	4.0	4.0	2.0	2.0
6	ALB	2.0	4.0	2.0	2.0
7	ALB	3.0	4.0	2.0	4.0
8	ALB	3.0	4.0	4.0	2.0
9	ALB	4.0	4.0	2.0	2.0
10	ALB	4.0	4.0	2.0	4.0
11	ALB	2.0	3.0	5.0	2.0
12	ALB	3.0	3.0	2.0	2.0
13	ALB	4.0	4.0	2.0	2.0
14	ALB	3.0	4.0	2.0	4.0
15	ALB	4.0	4.0	6.0	4.0

Fonte: Próprio Autor

Desse modo, foi calculado o percentual de cada tipo de resposta para cada uma das colunas. Esse procedimento foi executado por meio do seguinte comando:

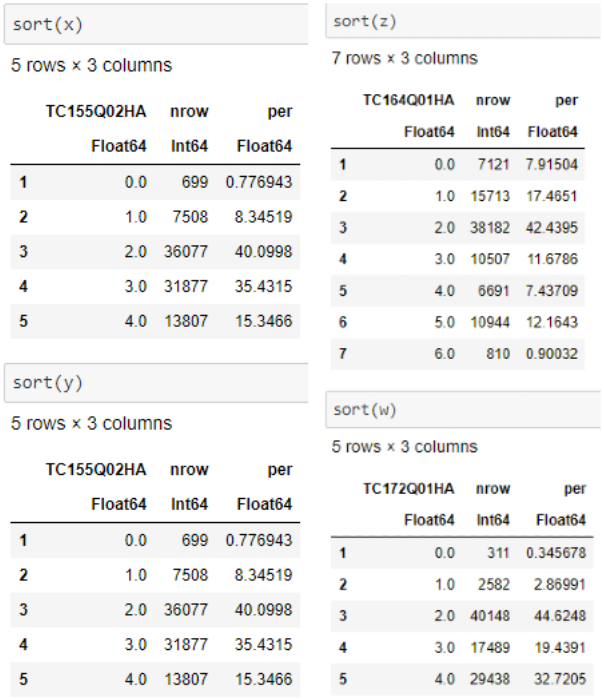
Figura 21 – Cálculo Percentual

```
t=nrow(q);
x = combine(nrow,groupby(q,:TC155Q02HA))
begin
    x[!, :per] = (x[!,:nrow] .* 100)/t;
    end
y = combine(nrow,groupby(q,:TC155Q02HA));
begin
    y[!, :per] = (y[!,:nrow] .* 100)/t;
    end
z = combine(nrow,groupby(q,:TC164Q01HA));
begin
    z[!, :per] = (z[!,:nrow] .* 100)/t;
    end/t;
w = combine(nrow,groupby(q,:TC172Q01HA));
begin
    w[!, :per] = (w[!,:nrow] .* 100)/t;
    end
```

Fonte: Próprio Autor

Os resultados encontrados pelo procedimento realizado na Figura 21 foram organizados da seguinte forma (Figura 22):

Figura 22 – Percentual dos tipos de respostas



```
sort(x)
```

5 rows × 3 columns

	TC155Q02HA	nrow	per
	Float64	Int64	Float64
1	0.0	699	0.776943
2	1.0	7508	8.34519
3	2.0	36077	40.0998
4	3.0	31877	35.4315
5	4.0	13807	15.3466

```
sort(y)
```

5 rows × 3 columns

	TC155Q02HA	nrow	per
	Float64	Int64	Float64
1	0.0	699	0.776943
2	1.0	7508	8.34519
3	2.0	36077	40.0998
4	3.0	31877	35.4315
5	4.0	13807	15.3466

```
sort(z)
```

7 rows × 3 columns

	TC164Q01HA	nrow	per
	Float64	Int64	Float64
1	0.0	7121	7.91504
2	1.0	15713	17.4651
3	2.0	38182	42.4395
4	3.0	10507	11.6786
5	4.0	6691	7.43709
6	5.0	10944	12.1643
7	6.0	810	0.90032

```
sort(w)
```

5 rows × 3 columns

	TC172Q01HA	nrow	per
	Float64	Int64	Float64
1	0.0	311	0.345678
2	1.0	2582	2.86991
3	2.0	40148	44.6248
4	3.0	17489	19.4391
5	4.0	29438	32.7205

Fonte: Próprio Autor

9 CONCLUSÃO

No decorrer deste trabalho, buscou-se apresentar e descrever conceitos, técnicas e ferramentas que estão atreladas à área de Ciência de Dados para, então, fosse possível uma exploração e manipulação de dados previamente selecionados. Por meio da linguagem Julia e os dados dos questionários de estudantes e professores do PISA 2018, foi possível exercer a maior parte dos procedimentos propostos e assim demonstrar a capacidade e enfatizar a importância dessa área do conhecimento.

A linguagem Julia mostrou-se adequada para aplicação dos conceitos dessa área, pois, ela possui bibliotecas com ferramentas simples e efetivas que auxiliam nas etapas de importação, seleção, filtragem e classificação quando se pretende trabalhar com um conjunto de dados desse tipo.

A essência deste trabalho foi, principalmente a demonstração de várias técnicas de filtragem e tratamento dos dados por meio da linguagem Julia. Na maioria das etapas, tanto os dados dos estudantes quanto o dos professores receberam tratamentos iguais ou similares.

Como a quantidade de dados gerados e armazenados aumentam constantemente, torna-se cada vez mais difícil para um ser humano visualizar um conjunto de dados no seu formato original, bem como obter informações que interessam a um indivíduo ou organização. Por meio de filtragens e manipulações corretas, um profissional da área de Ciência de Dados pode transformar um aglomerado de dados em um conjunto menor, de fácil visualização e com as características que mais lhe interessa, possibilitando assim, a busca por informações valiosas.

9.1 Para trabalhos futuros

A partir do que foi demonstrado nesse trabalho, é possível compreender que ainda existem muitas outras análises e aplicações que podem ser realizadas. Representações gráficas, por exemplo, poderiam demonstrar de forma visual comparações entre países, questões, alunos e professores. Além disso, a aplicação de algoritmos de aprendizagem de máquina podem ser realizados.

REFERÊNCIAS

AMARAL, Fernando. **Introdução à Ciência de Dados: mineração de dados e big data**. Alta Books, 2016.

ANGELONI, Maria Terezinha. Elementos intervenientes na tomada de decisão. – Universidade Metodista de São Paulo, 2003.

ASSUNÇÃO, Fernando. Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos. – UNIVERSIDADE DE SÃO PAULO, 2012.

BOBRIAKOV, Igor. **Comparison of top data science libraries for Python, R and Scala**. 2018. <https://medium.com/activewizards-machine-learning-company/comparison-of-top-data-science-libraries-for-python-r-and-scala-infographic-574069949267>. "acessado em 03/12/2019".

BORGES, Luiz Eduardo. **Python para desenvolvedores**. Novatec, 2014.

BUGNION, Pascal; MANIVANNAN, Arun; NICOLAS, Patrick R. **Scala: Guide for Data Science Professionals**. - Packt Publishing Ltd., 2017.

CAVALCANTE, Francisco; VIANNA, Fábio. **Regressão Linear Simples**. <http://www.cavalcanteassociados.com.br/utd/UpToDate151.pdf>. "acessado em 07/12/2019".

CHAPMAN, Stephan J. **Programação em MATLAB para engenheiros - Tradução 5ªEdição**. Cengage Learning, 2015.

CÔRTES, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. Mineração de Dados - Funcionalidades, Técnicas e abordagens. – PUC Rio, 2002.

COSTA, Antonio Alexandre Moura. Uma Abordagem Centrada na Filtragem Colaborativa para Redução do Custo Computacional do Método k-Nearest Neighbors. – Universidade Federal de Campina Grande Centro de Engenharia Eletrica e Informatica, 2014.

GARCIA, Simone Carboni. O uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde. – Universidade Federal do Rio Grande do Sul, 2003.

GRUS, Joel. **Data Science do zero: primeiras regras com o python**. Alta Books, 2016.

GUIDINI, Marilene Bertuol. APLICAÇÃO DO K-MEANSCLUSTER PARA CLASSIFICAR ESTILOS GERENCIAIS. – Revista Contemporânea de Economia e Gestão., 2008.

INDRUSIAK, Leonardo Soares. Linguagem Java. – JUG Rio Grande do Sul, 1996.

INEP. **RELATÓRIO BRASIL NO PISA 2018**. 2018. <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/pisa/resultados>. "acessado em 28/04/2021".

JUNIOR, Walter Teixeira Lima. Big Data, Jornalismo Computacional e Data Journalism: estrutura, pensamento e prática profissional na Web de dados. – Universidade Metodista de São Paulo, 2012.

JUPYTER. **Jupyter Notebook Documentation**. 2015. <https://jupyter-notebook.readthedocs.io/_/downloads/en/stable/pdf/>. "acessado em 29/04/2021".

LEITE, Renata Carvalho Macedo. Utilização de regressão logística simples na verificação da qualidade do ar atmosférico de Uberlândia. – Universidade Federal de Uberlândia, 2011.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de. Uma Introdução às Support Vector Machines. – Revista de Informática Teórica e Aplicada, 2007.

MAS, Jean-François. **Análisis espacial con R: Usa R como un Sistema de Información Geográfica**. European Scientific Institute, 2018.

MEHTA, Rajat. **Big Data Analytics with Java**. Packt Publishing Ltd., 2017.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. – Manole Ltda, 2003.

MORAES, Pedro Lee. Aprendizado por Reforço com Algoritmos Genéticos aplicado a Jogos. – PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO, 2019.

NETO, Cesare Di Girolamo. Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras. – INPE, São José dos Campos., 2014.

OECD. **PISA: Programme for International Student Assessment**. 2018. OECD Education Statistics (database), <https://doi.org/10.1787/data-00365-en>. "acessado em 17/10/2020".

OLIVEIRA, Paulo Felipe de; GUERRA, Saulo; MCDONNELL, Robert. **Ciência de dados com R**. - Instituto Brasileiro de Pesquisa e Análise de Dados, 2018.

PASSOS, Danielle Sandler dos. **BIG DATA, DATA SCIENCE E SEUS CONTRIBUTOS PARA O AVANÇO NO USO DA OPEN SOURCE INTELLIGENCE**. 2016. <http://www.revistasg.uff.br/index.php/sg/article/view/1026/524>. "acessado em 15/09/2019".

PEREIRA, João Marcello; SIQUEIRA, Mario Benjamim Baptista de. Linguagem de Programação Julia: Uma Alternativa open source e de alto desempenho ao MATLAB. – Revista Principia, 2016.

PISA. **PISA 2018 Technical Report**. 2018. <https://www.oecd.org/pisa/data/pisa2018technicalreport/>. "acessado em 29/04/2021".

PRICE, Jason. **Oracle Database 11G SQL: Domine SQL e PL/SQL no banco de dados Oracle**. Bookman, 2009.

RAUTENBERG, Sandro; CARMO, Paulo Ricardo Viviurka do. **Big Data e Ciência de Dados: complementaridade conceitual no processo de tomada de decisão**. Brazilian Journal of Information Studies, 2019.

VERONEZE, Rosana. Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla. – UNIVERSIDADE ESTADUAL DE CAMPINAS, 2011.

VIVIAN, Glaucio R.; CERVI, Cristiano R. Utilizando Técnicas de ´ Data Science para Definir o Perfil do Pesquisador Brasileiro da Área de Ciência da Computação. Instituto de Ciências Exatas e Geográficas ; Universidade de Passo Fundo.