Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

LEANDRO CÉSAR DA CRUZ

DATA SCIENCE:
DESENVOLVIMENTO DE APLICAÇÃO PARA ANÁLISE DE DADOS

Assis/SP

2018

Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

LEANDRO CÉSAR DA CRUZ

DATA SCIENCE:
DESENVOLVIMENTO DE APLICAÇÃO PARA ANÁLISE DE DADOS

Projeto de pesquisa apresentado ao curso de Bacharel em Ciência da Computação do Instituto Municipal de Ensino Superior de Assis – IMESA e a Fundação Educacional do Município de Assis – FEMA, como requisito parcial à obtenção do Certificado de Conclusão.
Linha de pesquisa: Ciências Exatas e da Terra.

Orientando(a): Leandro César da Cruz
Orientador(a): Prof. Dr. Luiz Ricardo Begosso

Assis/SP

2018

FICHA CATALOGRÁFICA

C957d CRUZ, Leandro César da

***Data Science*: Desenvolvimento de aplicação para análise de dados** / Leandro César da Cruz. Assis, 2018.

59 p.

Trabalho de Conclusão de Curso (Ciência da Computação) - Fundação Educacional do Município de Assis – FEMA

Orientador: Dr. Luiz Ricardo Begosso

1.*Data Science*. 2.Análise de Dados. 3.Tecnologia da Informação

CDD 005.113

DATA SCIENCE: DESENVOLVIMENTO DE APLICAÇÃO PARA ANÁLISE DE DADOS

LEANDRO CÉSAR DA CRUZ

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito do Curso de Graduação, avaliado pela seguinte comissão examinadora:

Orientador: _____
Prof. Dr. Luiz Ricardo Begosso

Examinador: _____
Prof. Me. Guilherme de Cleva Farto

Assis/SP
2018

DEDICATÓRIA

Dedico este trabalho a minha família, que sempre esteve ao meu lado ao longo destes quatro anos de estudo e dedicação, seja servindo de apoio ou me motivando a seguir em frente.

AGRADECIMENTO

Primeiramente agradeço a Deus por me dar conhecimento e sabedoria para chegar até aqui e por proporcionar esta grande oportunidade em minha vida.

Aos professores da Fundação Educacional do Município de Assis (FEMA/IMESA) pelos conhecimentos transmitidos em sala de aula, em especial ao orientador deste Trabalho de Conclusão de Curso Dr. Luiz Ricardo Begosso.

Aos amigos de sala, com quem convivi nos últimos quatro anos e por quem cultivo um grande sentimento de amizade e respeito, sobretudo aos amigos que contribuíram direta ou indiretamente na conclusão deste trabalho.

Por fim aos familiares, base pela qual sempre obtivi grande incentivo e que sempre me motivaram na busca de alcançar novos rumos.

Suba o primeiro degrau com fé. Não é necessário que você veja toda a escada. Apenas dê o primeiro passo.

Martin Luther King

RESUMO

Este trabalho visa conceituar o leitor acerca da abordagem conhecida como *Data Science*, que é definida como a área da computação responsável pela coleta de dados, análise de informações e tomada de decisões baseadas em elementos extraídos de um determinado banco de dados.

Por se tratar de um campo relativamente novo dentro da Tecnologia da Informação, este trabalho será de grande valia a futuros estudiosos no assunto. O projeto final, além do conhecimento teórico transmitido, destinar-se-á a criação de uma aplicação com Análise de Dados pela qual tornará possível visualizar de forma clara e objetiva o funcionamento de um sistema implementado para tal finalidade.

Palavra-chave: *Data Science*; Análise de Dados; Tecnologia da Informação.

ABSTRACT

This paper aims to conceptualize the reader about the approach known as Data Science, which define itself as an area of computation responsible for data collection, information analysis and decision-making based on elements collected by a particular database.
Because it is a relatively new field within Information Technology, this work will be of great value to future students in this area. The final project, besides transmitting theoretical knowledge, will be used to create an application with Data Analytics in which it will be possible to visualize in a clear and objective way the operation of a system implemented for this purpose.

Key words: Data science; Data Analytics; Information Technology.

LISTA DE ILUSTRAÇÕES

Figura 1 - Dados, informações, conhecimento e suas abstrações 19
Figura 2 - Competências para *Data Science* 24
Figura 3 - 3 V's de *Big Data* 27
Figura 4 – Fatores de aumento de dados 28
Figura 5 – Etapas da Análise de Dados 32
Figura 6 – Logomarca da linguagem de programação Python 36
Figura 7 – Logomarca do IPython 38
Figura 8 – Logomarca do Jupyter Notebook 38
Figura 9 – Estrutura do Jupyter Notebook 39
Figura 10 – Logomarca do Pandas 40
Figura 11 – Logomarca do NumPy 40
Figura 12 – Logomarca do Matplotlib 41
Figura 13 – Logomarca do Scikit-Learn 41
Figura 14 – Logomarca do Anaconda 42
Figura 15 – Adicionando os dados 45
Figura 16 – Aportuguesando os dados 46
Figura 17 – Dados nulos 47
Figura 18 – Dados duplicados 48
Figura 19 – Preenchimento de dados nulos 49
Figura 20 – Eliminação de dados duplicados 49
Figura 21 – Tumores maligno x benigno 50
Figura 22 – Estatística descritiva 51
Figura 23 – Impressão de histogramas 52
Figura 24 – Histogramas maligno x benigno 52
Figura 25 – Exemplo de *K-NN* com 4 vizinhos 53
Figura 26 – Definindo precisão com *K-NN* 54
Figura 27 – Taxa de risco 55

LISTA DE TABELAS

Tabela 1 – Visualização inicial do *Data Frame* 44
Tabela 2 – Visualização final do *Data Frame* 50

LISTA DE ABREVIATURAS E SIGLAS

API – Application Programming Interface

BI – Business Intelligence

CSV – Comma-Separated Values

HTML – HyperText Markup Language

JSON – JavaScript Object Notation

K-NN – K-Nearest Neighbor

ML – Machine Learning

TCC – Trabalho de Conclusão de Curso

TI -Tecnologia da Informação

SUMÁRIO

1 INTRODUÇÃO 16

1.1 OBJETIVOS 16

1.2 PÚBLICO ALVO 17

1.3 JUSTIFICATIVA 17

1.4 MOTIVAÇÃO 18

1.5 PERSPECTIVA DE CONTRIBUIÇÃO 18

1.6 METODOLOGIA 18

1.7 RECURSOS NECESSÁRIOS 19

1.8 ESTRUTURA DO TRABALHO 19

2 DADOS 20

2.1 DEFINIÇÃO DE DADOS 20

2.2 DADOS, INFORMAÇÕES E CONHECIMENTO 20

2.3 DADOS PARA *DATA SCIENCE* 22

3 *DATA SCIENCE* 24

3.1 INTRODUÇÃO A CIÊNCIA DE DADOS 24

3.1.1 Campos de atuação de *Data Science* 25

3.1.2 *Business Intelligence, Data Mining, Data Science e Data Analytics* 26

4 *BIG DATA* 28

4.1 CLASSIFICAÇÃO DE *BIG DATA* 28

4.2 EXPANSÃO DO *BIG DATA* 29

4.3 CONCLUSÃO DE *BIG DATA* 30

5 *MACHINE LEARNING*..... 32

5.1 CONCLUSÃO DE *MACHINE LEARNING*.....32

6 ANÁLISE DE DADOS 33

6.1 QUESTIONAMENTO.....33

6.2 PREPARAÇÃO.....33

6.3 EXPLORAÇÃO 34

6.4 CONCLUSÃO 34

6.5 COMUNICAÇÃO.....34

6.6 AMOSTRA DE ANÁLISE DE DADOS 35

7 FERRAMENTAS..... 36

7.1 LINGUAGEM DE PROGRAMAÇÃO 36

7.2 PACOTES PYTHON PARA ANÁLISE DE DADOS.....37

7.2.1 IPython 37

7.2.2 Jupyter Notebook.....38

7.2.3 Pandas 39

7.2.4 NumPy 40

7.2.5 Matplotlib.....40

7.2.1 Scikit-Learn.....41

7.2.1 Anaconda.....42

8 ESTUDO DE CASO 43

8.1 VISÃO GERAL DO CASO 43

8.2 ANÁLISE DO CASO 43

8.2.1 Iniciando a análise 43
8.2.2 Questionar 44
8.2.3 Preparar 45
8.2.4 Explorar 50
8.2.5 Concluir 52
8.2.6 Comunicar 54
9 CONCLUSÃO 55
9.1 CONSIDERAÇÕES FINAIS 56
9.2 PROJETOS FUTUROS 56
REFERÊNCIAS 57

1 INTRODUÇÃO

Desde o fim da segunda guerra mundial, o mundo vive uma constante evolução no que tange o desenvolvimento de novas tecnologias. Com o advento da rede mundial de computadores – no final dos anos 60 – o volume derivado da troca de informações entre pessoas cresceu de forma exponencial; gerando assim um grande gargalo em relação ao modo mais eficiente de se fazer a manipulação e o processamento destes dados. Com foco nesta problemática, engenheiros e cientistas da computação vêm ao longo do tempo criando diversas técnicas, que se complementam na busca de um método único de implementação que satisfaça tal necessidade.

Como um dos processos mais importantes disponíveis atualmente, a Ciência de Dados - ou *Data Science* como é mais conhecida - foi desenvolvida com a finalidade de suprir esta lacuna computacional, gerando assim soluções e ideias a partir de fontes de dados distintas. Sua versatilidade permite aplicá-la em diversas áreas dentro e fora da computação, tais como: reconhecimento de imagem, Análise de Dados, inteligência artificial, *Big Data*, *Machine Learning*, *Data Mining*, robótica, negócios, entre outras...

O foco central deste trabalho é definir um embasamento teórico no intuito de desenvolver uma ferramenta que demonstre pragmaticamente a implementação de *Data Science* na análise dos dados coletados.

1.1 OBJETIVOS

Este Trabalho de Conclusão de Curso (TCC) objetiva apresentar as especificações e a implementação de sistemas computacionais que comprovem a real eficácia de *Data Science* na busca de extrair ideias e informações relevantes, que possam ser analisadas e transformadas ao ponto de gerarem resultados positivos dentro de uma determinada organização.

Apesar de se tratar de um termo presente já há algum tempo no mercado de Tecnologia da Informação (TI), a Ciência de Dados pode ser caracterizada como uma metodologia relativamente nova, haja vista que seu real emprego só se dá em áreas muito restritas e específicas; o que nos remete a outro propósito crucial deste trabalho, que é o de demostrar sua validação em vários ramos onde a computação se aplica.

Conforme deseja-se demonstrar, “dados e a capacidade de extrair conhecimento útil dos mesmos devem ser considerados os principais ativos estratégicos fundamentais”

(FAWCETT; PROVOST, 2013, p. 9). Sendo assim podem ser utilizados em qualquer setor onde se necessite de algum tipo de estratégia organizacional que busque a otimização de tempo, espaço, esforço ou recursos.

1.2 PÚBLICO ALVO

O público alvo, quando se trata de *Data Science,* se refere a qualquer ramo de atividade onde a computação possa ser aplicada e que possua algum tipo de fonte de dados. Portanto pode ser inserida em praticamente todas as áreas onde se gera algum tipo de informação.

Este trabalho tem como foco principal o estudo e a implementação de *Data Science* na forma de demonstração do funcionamento da análise dos dados na área da saúde.

1.3 JUSTIFICATIVA

Segundo PIERSON (2017, p. 23), “muitas pessoas acreditam que apenas grandes organizações que tenham financiamento maciço estão implementando metodologias de *Data Science* para otimizar e melhorar seus negócios, mas a realidade não é bem assim. A ploriferação de dados criou uma demanda de *Insight*, e essa demanda está incorporada em muitos aspectos de nossa cultura moderna”. É justamente neste contexto que este projeto se apoia, pois, a utilização de *Data Science* se tornou atualmente uma necessidade e pode ser utilizado em grande escala até mesmo por pequenas corporações ou orgãos governamentais.

O desenvolvimento deste projeto não se justifica apenas em demonstrar o funcionamento da implementação, mas também comprovar sua importância na busca de um bem-estar coletivo dentro de uma determinada comunidade. A Ciência de Dados pode gerar resultados significativos tanto no setor público como no privado. No setor público pode agir em áreas cruciais das sociedades - sobretudo a brasileira – como por exemplo: fiscalização, saúde, segurança, educação, trânsito e transporte; e no setor privado buscar a melhoria no fluxo de negócios de campos de atuações variados, tais como: agricultura, fábricas, indústrias, empresas de produtos, empresas de serviços, entre outros... Ajudando assim a aumentar os lucros ou evitar grandes perdas.

1.4 MOTIVAÇÃO

A principal razão para desenvolvimento deste trabalho se deu pelo fato de que atualmente o setor da tecnologia conhecida como *Data Science,* mesmo que famigerada por pessoas do ramo, pode ser considerada uma tecnologia de nicho. Hoje basicamente empresas de grande porte oferecem este tipo de serviço, o que o torna de maneira geral praticamente inacessível a empresas de pequeno e médio porte (devido ao elevado custo) e a órgãos governamentais (por conta de fatores burocráticos).

Devido a sua capacidade preditiva, a Análise de Dados deixou de ser um diferencial e passou a ser algo fundamental no processo de gestão. Por tornar possível uma previsão a longo prazo de fatos que possam vir a ocorrer, a Ciência de Dados possibilita a gerentes, diretores e administradores uma visão futura de problemas e possíveis soluções; auxiliando assim no processo de tomada de decisão e evitando que erros aconteçam.

1.5 PERSPECTIVA DE CONTRIBUIÇÃO

Ao final do projeto com o término da implementação e apoiado na teoria levantada, espera-se poder demonstrar o quanto a democratização da *Data Science* se tornou algo indispensável para as organizações. Espera-se também certificar que sua aplicação não é algo tão distante da realidade tanto quanto se imagina.

A contribuição maior deste trabalho estará contida em todo o material de acesso do mesmo, que servirá de base para projetos futuros de pesquisas e artigos que venham a ser realizados nesta área, sobretudo no âmbito da computação.

1.6 METODOLOGIA

O andamento simulado em relação a análise dos dados será feito com base em dados coletados de site especializado em *Data Set*. O trabalho se auxiliará de algumas ferramentas de programação tais como Python (linguagem de programação), que hoje é uma das linguagens mais indicadas para Análise de Dados devido a vários recursos para este fim já embutidos na linguagem. Como paradigma de programação será utilizada a Orientação a Objetos, por oferecer recursos indispensáveis ao desenvolvimento focado em Ciência de Dados.

Como ponto de partida serão colhidos dados reais acerca do câncer de mama, disponibilizados pela Universidade de *Wisconsin,* que servirão de apoio ao avanço na análise dos dados.

Para uma correta documentação o sistema será modelado utilizando a ferramenta Jupyter Notebook, que se trata de um ambiente criado para integrar em um só lugar, diversos recursos de programação mesclados com sua devida documentação.

1.7 RECURSOS NECESSÁRIOS

Hardware: Ultrabook Acer Aspire M; processador Intel Core i5 1.8 GHz; disco rígido 500 GB, memória DDR 3 de 6 MB.

Sistema Operational: Windows 10 Pro.

Software: IPython.

Linguagem de programação: Python 3.

Pacotes para Análise de Dados: IPython, Jupyter Notebook, Pandas, Numpy, Matplotlib, Scikit-Learn, Anaconda.

1.8 ESTRUTURA DO TRABALHO

- **Capítulo 1 – Introdução:** Neste capítulo será apresentado os conceitos primordiais de *Data Science* e seu ramo de atuação.
- **Capítulo 2 – Dados:** Uma abordagem acerca de dados, explicando sua importância ao tema e sua diferença em relação a informação e conhecimento.
- **Capítulo 3 – *Data Science*:** O capítulo 3 vem acompanhado de uma breve introdução de *Data Science,* junto a explicação de sua necessidade atualmente.
- **Capítulo 4 – *Big Data*:** Esclarecimento do *Big Data*, que pode estar presente quando se trabalha com um volume muito grande de dados.
- **Capítulo 5 – *Machine Learning*:** A importância da Aprendizagem de Máquina em *Data Science*.
- **Capítulo 6 – Análise de Dados:** Análise dos dados com questionamento, preparo, exploração, conclusão e comunicação dos resultados encontrados.
- **Capítulo 7 – Ferramentas:** Ferramentas necessárias para a implementação do projeto.
- **Capítulo 8 – Estudo de Caso:** Sistema a ser implementado no decorrer do projeto.
- **Capítulo 9 – Conclusão:** Esclarecimento final acerca dos resultados coletados pelo projeto.
- **Referências**

2 DADOS

Antes mesmo do conteúdo relacionado a *Data Science*, este trabalho concentrou-se no entendimento dos dados, que constituem um dos elementos primordiais na compreensão deste novo objeto de estudos.

Por mais que a Ciência de Dados seja um recurso relativamente recente e que imprime conceitos que estão estritamente ligados ao conhecimento de ferramentas e técnicas computacionais modernas, o estudo de sua implementação parte do pressuposto de que o leitor já obtenha o entendimento da importância dos dados dentro do escopo proposto em Análise de Dados. Portanto os dados tendem a ser o ponto de partida a qualquer pesquisa destinada a este campo do conhecimento humano.

Tornar claro o termo “dados”, assim como demostrar as diferenças entre dados, informações e conhecimento são tarefas indispensáveis para um futuro avanço no entendimento de Ciência de Dados, já que no decorrer desta pesquisa estes recursos serão abordados de formas diversas.

2.1 DEFINIÇÃO DE DADOS

Buscando em sua origem, o vocábulo “dados” tem suas raízes na palavra latina *datu* que - em uma tradução livre - corresponde a palavra aportuguesada “dar”. A partir daí pode-se deduzir que os dados são na realidade fatos dados, pelos quais podemos deduzir fatos adicionais.” (DATE, 2003, p. 13). Portanto representando em uma avaliação mais completa, podemos definir os dados como um conjunto de fatos ou valores quantitativos, em estado bruto, que são dados na forma de medição ou adquiridos por meio de alguma observação previamente documentada e que possui a finalidade de obter alguma informação posterior.

Por conta de sua característica quantitativa SETZER (1999, p. 2) exalta que “...um dado é necessariamente uma unidade matemática e, desta forma, puramente sintática”, ou seja, os dados podem facilmente ser armazenados e processados por qualquer sistema computacional destinado para tal finalidade.

2.2 DADOS, INFORMAÇÕES E CONHECIMENTO

Apesar de sua grande importância em *Data Science*, os dados brutos representam apenas o estágio inicial no processo de geração de resultados, e podem ser considerados

como a matéria-prima no processo de análise dos mesmos. Com esta definição em mente “partimos do princípio de que os dados são a fonte predominante para a obtenção de informação” (BRAGA, 2005, p. 16).

Antes de gerar um resultado final os dados percorrem diversos processos, passando de meros dados a informações e de informações a conhecimentos úteis; que são colhidos na procura da obtenção de *insights*.

Na medida em nos aprofundamos nas disparidades existentes entre dados, informações e conhecimento, também aumentamos o nível de abstração das expressões estudadas. Conforme demonstra a figura abaixo, quanto mais um dado isolado passa a ser utilizado como informação, este mesmo passa também a ter uma maior abstração em relação ao seu significado e sua complexidade; assim como também aumenta seu grau de abstração na transição de informação em conhecimento.

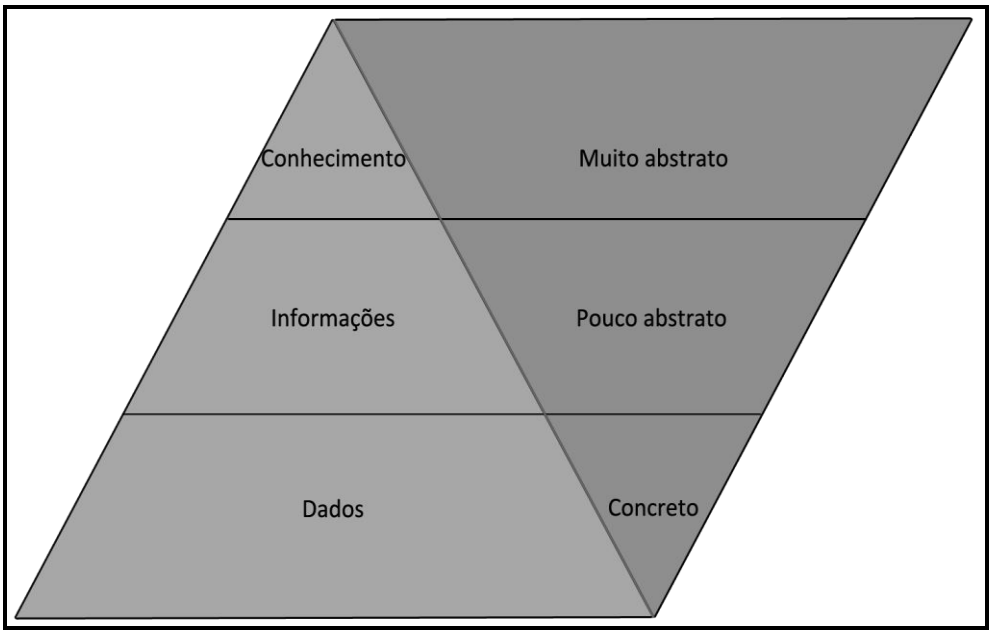

Figura 1: Dados, informações, conhecimento e suas abstrações (THIERAUF, 1999, p. 23)

De acordo com THIERAUF (1999, p. 6-7) os dados são o ponto mais baixo, uma coleção não estruturada de fatos e números; já a informação é o próximo nível, e é considerada como dados mais estruturados.

O acesso aos dados, acarretou em uma maior liberdade de desenvolver novas abstrações ao ponto de transformá-los em informações e posteriormente em conhecimento. Através

do conhecimento foi possível alcançar a geração de resultados na forma de sabedorias ou *insights*.

O uso dos dados pode ser definido como uma unidade exclusivamente utilizada por uma máquina ou neste caso, por um computador. Este computador possui uma linha de pensamento exata – sentido concreto - e de baixa abstração em relação ao que estes dados realmente significam em termos mais filosóficos.

As informações e seus sucessores, por sua vez, contam com uma maior quantidade de abstração em relação ao conteúdo analisado, onde, para que possam ser utilizadas é preciso um certo grau de abstração. Esta abstração deve transmitir algo que possua significado por meio da interpretação humana.

Podemos dizer que não é possível processar informações diretamente por um computador. Para isso é necessário reduzi-las a dados (SETZER, 1999, p. 2-3).

Os dados representam uma informação armazenada e não deve ser considerada uma informação propriamente dita, mas sim sua representação em forma de número. Podemos dizer que os computadores trabalham com dados e os humanos com informações.

Feito uma análise mais aprofundada o conhecimento transcende a complexidade da informação, na medida que depende de um conjunto de características e experiência do indivíduo que o detém, ou seja, o conhecimento depende de um *Know-How* particular intrínseco ao indivíduo que o possui.

Estes conceitos são importantes porque é justamente a partir deles que o cientista de dados toma suas decisões no intuito de gerar *insights* importantes retirados de suas análises. Quanto mais conhecimento se tiver do objeto de estudo analisado no sistema de *Data Science*, maior será a capacidade de gerar resultados satisfatórios.

2.3 DADOS PARA DATA SCIENCE

Como a Ciência de Dados trabalha fundamentada em conhecimentos gerados na forma de *insights,* a partir da análise das informações processadas, é fácil concluir que quanto maior a quantidade de dados disponíveis, maior serão as informações obtidas; o que por sua vez agregará numa melhor capacidade de absorção de conhecimento para a geração de ideias. Por este motivo é de extrema importância que trabalhemos com grandes volumes de dados. Na Ciência da Computação um aglomerado de grande quantidade de dados armazenados e processados, pronto para ser utilizado leva o nome de *Big Data*.

Por conta de sua grande relevância dentro de nossos estudos o assunto *Big Data* será melhor abordado em um capítulo posterior, porém entende-se que em Ciência de Dados seu uso não é obrigatório, contudo acarreta em resultados mais satisfatórios.

Os tipos de dados que são utilizados em *Data Science* são distintos, o que permite a utilização tanto dado estruturado, como não estruturado.

Dados estruturados: São dados com uma organização padrão estabelecida para uma melhor organização e recuperação dos mesmos. É organizado na forma de tabelas (linhas e colunas), sendo assim, o tipo mais comum de se encontrar em aplicações cujo uso do volume de informações é considerado baixo ou médio.

Dados semi-estruturados: Um intermédio entre dados estruturados e não estruturados. Nele os dados não são estritamente tipados e são coletados de maneira casual. Não possui uma estrutura padrão para as informações coletadas. Seus atributos são presentes apenas em algumas entidades.

Dados não estruturados: registra o maior número de informações possíveis que se deseja armazenar, mas sem se preocupar com suas estruturas. É utilizado em larga escala para armazenar tipos distintos de dados como formato de textos, vídeos, imagens, entre outros. “As páginas web em *HTML* que contêm alguns dados são considerados dados não estruturados” (ELMASRI; NAVATHE, 2011, p. 281).

3 DATA SCIENCE

A rápida evolução da Tecnologia da Informação derivada do surgimento da rede mundial de computadores, passando pela diminuição do valor de mercado dos microcomputadores e pela criação de diversos dispositivos móveis; trouxe consigo uma quantidade absurda de dados compartilhados entre pessoas e máquinas.

A união destes dados embora muito importante para a evolução da sociedade, pouco nos oferece em termos de análise preditiva, já que isolados, os dados não passam de um aglomerado de números armazenados por um sistema gerenciador de banco de dados e que são acessados pelo usuário comum apenas como meio de obter algum tipo de informação.

Foi justamente pela necessidade de se extrair um conteúdo mais relevante destes dados que a Ciência de Dados foi criada. Sua tarefa é a de localizar padrões, para assim criar soluções a partir de informações que possam ser aplicadas em forma de conhecimento, afim de produzir novas ideias sobre um determinado objeto de estudo.

O termo designado para firmar o nome desta ciência não se encontra bem definido entre os escritores. Mesmo pesquisando em diversas fontes, percebe-se que muitos fazem uso do termo em inglês "*Data Science*" enquanto outros preferem a definição aportuguesada "Ciência de Dados". Este documento – como o próprio tema indica – tende a ter uma inclinação a primeira forma, por conta de sua maior ocorrência ao longo deste estudo.

3.1 INTRODUÇÃO A CIÊNCIA DE DADOS

O primeiro contato com a Ciência de Dados para quem se inicia em um estudo mais engajado no assunto gera um certo desconforto. Isso por conta do caráter multidisciplinar muitas vezes encontrado em suas documentações.

Sua multidisciplinaridade se deve ao fato de que *Data Science* por se tratar de uma ciência, aborda diversas áreas do conhecimento; áreas estas que sofrem a intersecção de temas correlatos gerando assim um resultado final transmitido pelo cientista de dados.

Como forma de entendimento, campos de atuações variados que fazem parte deste processo podem ser citados, tais como: banco de dados, programação de computadores, *Machine Learning*, *Data Mining*, Análise de Dados, estatística, matemática, física, engenharia de *software*, negócios, entre outros. Outro ponto importante a ser observado para novos estudiosos no assunto é sua pouca documentação, já que se trata de um

tema totalmente inovador para os tempos atuais (2018), na medida que ainda não é utilizado de forma efetiva no mercado.

Alberto Boschetti, Luca Massaron (2016, p. 8)

> A Ciência de Dados é um domínio de conhecimento relativamente novo, embora seus componentes principais tenham sido estudados e pesquisados por muitos anos pela comunidade de Ciência da Computação. Seus componentes incluem álgebra linear, modelagem estatística, visualização, linguística corporal, analise de gráficos, aprendizado de máquina, inteligência de negócios, armazenamento e recuperação de dados.

O volume absurdo de dados que são gerados individualmente e que crescem de forma variada com uma velocidade gradativa ao longo dos tempos, é chamado de *Big Data*. Buscar estes dados e "torturá-los", ao ponto de extrair informações menores e úteis é o foco principal da Ciência de Dados. Pode-se dizer também que *Data Science* é a forma mais efeciente desenvolvida até o momento de prever o futuro baseado em um olhar humano dos dados. No meio deste processo todo, o cientista de dados é o profissional que deve agir com tomada de decisões a partir de informações geradas através dos dados coletados, mesclando-os com seus conhecimentos individuais adquiridos ao longo do tempo.

3.1.1 Campos de atuação de *Data Science*

Como foi exposto no início do tópico anterior, a dificuldade encontrada para definir em poucas palavras a que se refere a tecnologia *Data Science*, se deve ao fato de que esta pertence a uma ciência multidisciplinar que incorpora múltiplas áreas do conhecimento humano.

Para se alcançar um resultado final satisfatório com a Ciência de Dados é de fundamental importância a união de três competências básicas, que são: conhecimento computacional (*Computer Science* ou *Hacking Skills*), conhecimento exato (*Math, Statistics and knowledge*) e conhecimento de especialista (*Subject Matter Expertise*).

Segundo VANDERPLAS (2016, p. 11) *Data Science* é "o conjunto de habilidades interdisciplinares que estão se tornando cada vez mais importantes em muitas aplicações em toda a indústria e academia."

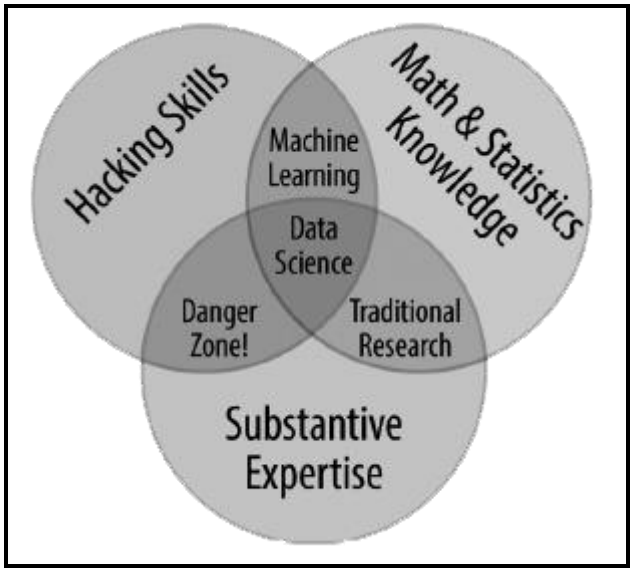

Figura 2: Competências para *Data Science* (VANDERPLAS, 2016, p. 11)

Por conta de seu vasto leque de multidisciplinaridade foi criado o termo em inglês "*Unicorn Data Scientist*" ou "Cientista de Dados Unicórnio", que se designa ao cientista de dados que possui todas as competências necessárias para se trabalhar com Ciência de Dados. A palavra unicórnio se deve ao fato de que se trata de um animal fictício ou inexistente; o mesmo conceito de ficção que se aplica ao profissional de *Data Science* que domine individualmente todos os graus de competência nesta ciência.

Além de sua característica multidisciplinar a Ciência de Dados possui também a qualidade de atuar em multiplas áreas, já que pode ser aplicada em quaisquer domínios que gerem uma quantidade substancial de dados.

3.1.2 *Business Intelligence, Data Mining, Data Science* e *Data Analytics*

Muitas dúvidas surgem em torno do termo *Data Science* em relação a outros termos que embora pareçam sinônimos, possuem formas de atuações totalmente diversas. As dúvidas mais comuns surgem a partir das expressões: *Business Intelligence, Data Mining, Data Science* e *Data Analytics*. Por conta disso é também muito importante saber algumas diferenças entre estas áreas, para que não haja dúvidas posteriores em relação aos devidos termos.

Business Intelligence (BI): analisa fatos que já tenha ocorrido em um determinado momento se fundamentando em dados exatos que já existam, não se importando tanto

quanto em *Data Science* em realizar predições em prazos longínquos. Em *BI* o trabalho é organizado em cima do que está acontecendo no momento (médio e curto prazo) o que faz com que tenha de tomar decisões mais pontuais. Um exemplo seria o diagnóstico de vendas dentro de um determinado mês em uma *e-commerce*; onde todo o trabalho seria feito sobre dados existentes e com um foco menor de previsão quando se compara a Ciência de Dados.

Data Mining: Data Mining ou Mineração de Dados explora os dados em busca de padrões utilizando técnicas analíticas guiadas por uma máquina. Nela os resultados são validados a partir de novos subconjuntos de dados chegando em objetivo final quando gerado um prognóstico pré-estabelecido. Pode-se dizer que faz o uso da computação quase que em sua totalidade, na construção de seus processos de trabalho.

Data Science: Assim como *Data Mining,* é preditiva, porém trabalha com os dados utilizando-os como informações e conhecimento de especialistas. A diferença principal se dá pelo fato de que esta tecnologia atua com o adjunto de técnicas científicas mais variadas, tais como: Estatística, *Machine Learning*, *Data Analytics*, *Data Mining* entre outras...

Data Analytics: Também conhecida como Análises de Dados é o processo pelo qual procura-se inspecionar, limpar, transformar e modelar dados. “É geralmente vista como um componente da Ciência de Dados. Usada para entender como são os dados de uma organização a Ciência de Dados usa Analytics para resolver problemas (OLAVSRUD, 2018). Devido sua grande importância neste projeto este assunto terá um tópico específico.

4 BIG DATA

Como já exposto no tópico anterior um dos recursos padrões em Ciência de Dados é a utilização de *Big Data*, pela necessidade de se trabalhar com um volume substancialmente grande de dados, a fim de otimizar as informações a serem analisadas. "*Big Data* é um termo geral para qualquer coleção de dados tão grande ou complexo que torna difícil seu processamento, utilizando o gerenciamento tradicional de dados". (CIELEN; MEYSMAN; MOHAMED, 2016, p. 1).

Apesar de ter uma característica que permite a manipulação e a estruturação de um grande volume de dados, o *Big Data* não se limita única e exclusivamente a isto. A seguir para um melhor entendimento do assunto novos adjetivos importantes serão atribuídos.

4.1 CLASSIFICAÇÃO DE BIG DATA

Como podemos concluir de forma intuitíva, o termo *Big Data* sugere uma referência que está diretamente relacionada a uma dimensão sem precedentes em relação ao volume dos dados. Este fato certamente se confirma, porém devemos ter ciência de que seus fundamentos não se limitam única e exclusivamente a esta definição.

Um dos termos mais populares por estudiosos de *Big Data* é a propriedade dos "3 V's", que enumeram as 3 principais bases deste mecanismo, que são: Volume, Variedade e Velocidade.

Volume: Faz referência ao grande volume de dados e/ou informações que são utilizados. Não obstante, esta sem dúvida é a característica mais significatica dentro de um ambiente de *Big Data*.

Definir uma determinada quantidade de dados como sendo *Big Data* não é algo simples. O indicador primordial em apontar em que momento um certo volume de dados se torna ou não *Big Data,* é a quantidade de dados em relação ao ponto de atuação específico onde estes se aplicam. O volume de dados que é classificado como *Big Data* em uma empresa aeroespacial é diderente em relação a uma empresa de e-commerce ou uma empresa de entretenimento por exemplo.

Um ponto que nos mostra como não é possível definir um número exato que se encaixe em *Big Data* é o fato de que a tentativa exaustiva de definir algo como *Big Data* não seria muito útil porque os conjuntos de dados estão crescendo a cada ano (WILLIAMSON,

2015, p. 8-9). Os dados que hoje são medidos como grande em relação a capacidade de espaço de armazenamento, podem deixar de ser daqui a 10 anos por exemplo.

Variedade: *Big Data* não faz uso apenas de dados estruturados, mas também semi-estruturados e não estruturados. Seu significado também se aplica em relação aos locais onde o *Big Data* deve ser aplicado, passando por organizações de filosofias e estruturas completamente diferentes como: agro-indústria, gestão pública, comércio varejista, comércio atacadista, pesquisas científicas, et cetera...

Velocidade: Diz respeito a velocidade de utilização destes dados; em *Big Data* como o volume de dados é substancialmente maior, existe a exigência de que os mesmos sejam coletados e analisados de forma mais rápida. (MARQUESONE, 2016, p. 01-09)

Os serviços de *streaming* que transmitem e processam milhares de dados em tempo real fazem parte de um grupo de tecnologias que só existem por conta desta característica de velocidade aqui encontrada.

A seguir (Figura 3) temos uma imagem contendo os 3 V's assim como algumas das características principais de cada um deles:

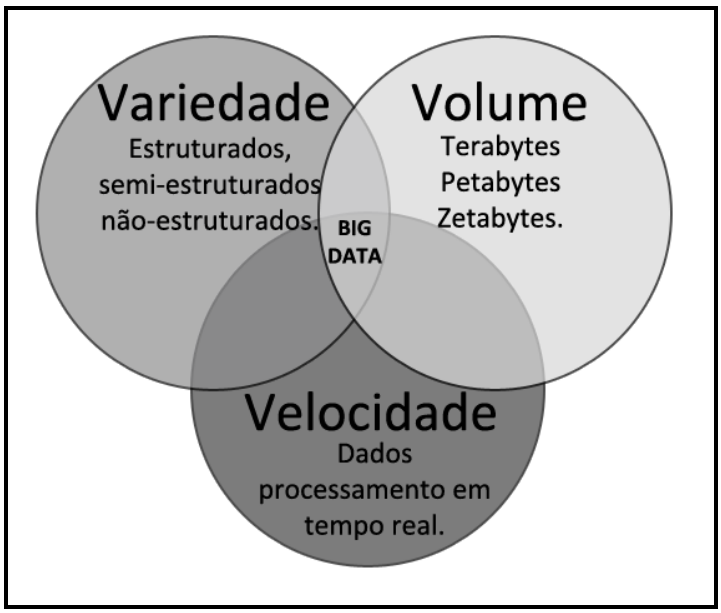

Figura 3: 3 V's de *Big Data* (MARQUESONE, 2016, p. 8)

4.2 EXPANSÃO DO BIG DATA

A rápida evolução do mercado de TI fez com que o desenvolvimento de *Big Data* se tornasse cada vez mais acelerado e necessário.

Alguns fatores se colidiram e ajudaram a multiplicar de forma exponencial a quantidade de dados disponíveis em nosso planeta.

O grande aumento de dados disponíveis para consulta e manipulação, sem dúvida se deu com o grande “boom” da internet no final dos anos 60. Sua chegada tornou o mundo mais coeso e informatizado, fato que foi essencial para firmar a importância dos microcomputadores na vida das pessoas. Outra grande revolução se deu com o surgimento de dispositivos móveis, que trouxe consigo uma maior facilidade no compartilhamento de informações. A arquitetura de computadores também teve um papel importante, já que, com a passar dos anos, produtos foram se tornando cada vez mais baratos e com uma capacidade de processamento mais eficiente.

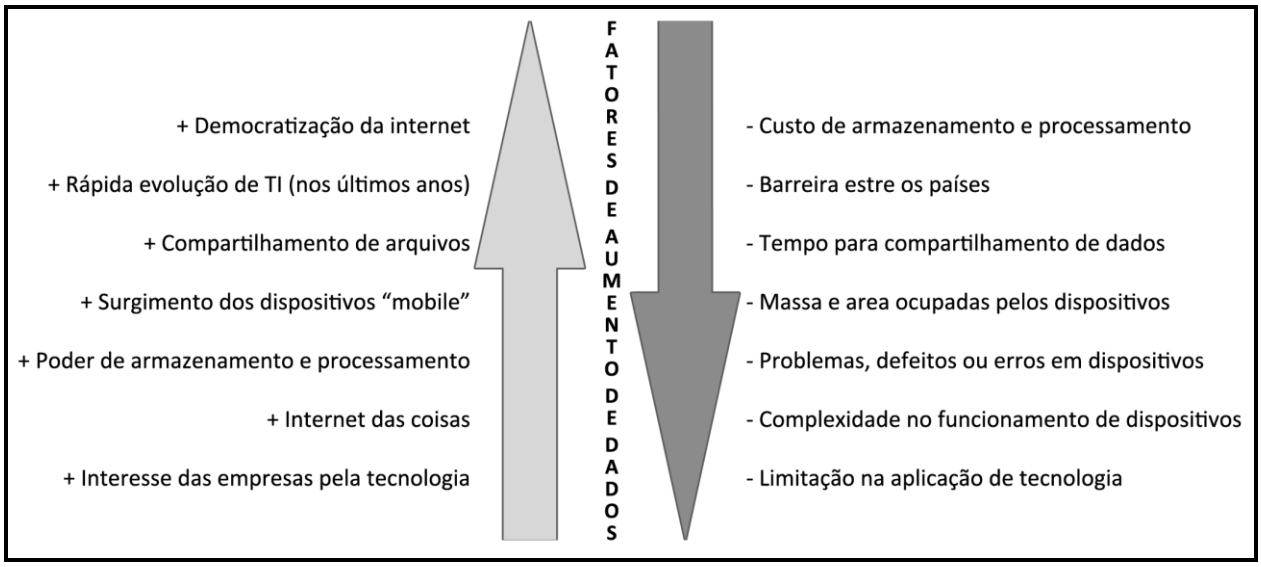

Figura 4: Fatores de aumento de dados (MARQUESONE, 2016, p. 5)

Entre estes fatores centenas de outros se destacam e firmam assim a certeza da importância de se trabalhar com *Big Data.* Como *Big Data* não é o foco deste projeto nos limitaremos apenas a estas definições.

4.3 CONCLUSÃO DE BIG DATA

Por conta de sua grande versatilidade e devido aos benefícios trazidos com seu uso, conclui-se que o ideal em se trabalhar com Ciência de Dados é incluir na aplicação o *Big Data*, afim de retornar informações mais confiáveis. O uso do *Big Data* auxilia no fato de que muitos dados coletados devolvem um resultado com um percentual mais elevado em

relação a precisão nas manipulações futuras. Todavia mesmo com todas as suas vantagens, devemos destacar que o uso de *Big Data* neste caso não se trata de regra obrigatória, sendo então possível a análise dos dados que não se encaixem nos conceitos de *Big Data* citados acima; mesmo que a fonte de consulta não seja tão assertiva quanto.

5 MACHINE LEARNING

O que automatiza a coleta de informações e maximiza a velocidade em que estas são fornecidas é o *Machine Learning* (*ML*). Seu papel é o de ensinar o computador a concluir suas tarefas sozinho, na medida em que se obtém uma grande sequência de dados para manipulação. O algoritmo implementado com Aprendizagem de Máquina ajuda na extração de ideias, por maximizar o processo da análise dos dados brutos. O objetivo da *ML* é o de ensinar a máquina (ou *software*) a realizar tarefas, fornecendo-lhe alguns exemplos (RICHERT; COELHO, 2013, p. 7-9).

Como bem definido pela citação de Mike Roberts no livro *Introducing Data Science*: “O aprendizado de máquina é o processo pelo qual um computador pode trabalhar com mais precisão à medida que recolhe e aprende com os dados fornecidos” (CIELEN; MEYSMAN; MOHAMED, 2016, p. 58). Em *Data Science* esse recurso pode deixar explícito informações que passariam desapercebidas pelo olhar humano ao executar a análise dos dados.

5.1 CONCLUSÃO DE MACHINE LEARNING

Conclui-se que o uso da Aprendizagem de Máquina é o elemento principal que difere um processo de Análise de Dados simples para uma Análise de dados em *Data Science*. A análise dos dados não necessariamente necessita de um computador para ser executada com sucesso, entretanto se o objetivo for automatizar processos pelo uso de *ML* o elemento máquina é indispensável.

6 ANÁLISE DE DADOS

Em todo o procedimento de Ciência de Dados, a análise corresponde a um dos componentes mais importantes. É nela que o material humano age de forma mais intensa e relevante, pelo intermédio do analista de dados.

O analista de dados é o indivíduo que trabalha com dados brutos, fazendo perguntas a respeito de um determinado tema. Ele não necessariamente depende da Tecnologia da Informação para validar seu trabalho, porém sabe-se que os recursos computacionais potencializam imensuravelmente o processo de pesquisa. Os computadores analisam os dados e buscam critérios que se juntam com os objetivos estabelecidos pelos humanos (WEIS, 1999, p. 32).

Como o fluxo da Análise de Dados é constituído de diversas etapas, sua explicação neste trabalho foi subdividida entre seus 5 principais tópicos que são: questionamento, preparação, exploração, conclusão e comunicação.

6.1 QUESTIONAMENTO

O passo inicial na análise de dados é questionar sobre o problema a ser solucionado e definir qual parte dos dados serão mais importantes. Ao questionar, o cenário se encontrará desenhado em torno de uma entre duas hipóteses. Ou se inicia o questionamento tendo os dados disponíveis; ou questiona-se primeiro para depois iniciar a coleta dos dados. Em ambos os casos uma boa pergunta ajudará a se concentrar nos aspectos mais relevantes e a direcionar a análise para *insights* significativos. (LEE; Udacity, 2017)

6.2 PREPARAÇÃO

Na preparação os dados são obtidos de forma que possibilite trabalhar em três etapas: reunir, avaliar e limpar. Primeiramente reúnem-se os dados relevantes para responder as perguntas do questionamento. Em seguida avaliam-se os dados para encontrar qualquer problema em relação a qualidade ou estrutura dos mesmos. Por fim os dados são limpos, removendo ou substituindo aqueles que estiverem fora do padrão, na garantia de que o conjunto final seja da mais alta qualidade e bem estruturado possível. (LEE; Udacity, 2017)

6.3 EXPLORAÇÃO

A análise exploratória dos dados, também identificada pela sigla EDA (*Exploratory Data Analytics*) se baseia em encontrar padrões nos dados e extrair intuições sobre o assunto em que se está trabalhando. Após explorar pode-se realizar a engenharia de recursos em que se removem *outlier* (dados que se diferenciam drasticamente), criando assim melhores recursos com os dados. (LEE; Udacity, 2017)

6.4 CONCLUSÃO

Tirar conclusões é o passo final antes de comunicar os resultados finais obtidos, nele pode-se até conseguir identificar previsões de eventos. Por conta de seu caráter preditivo a conclusão é realizada com a implementação de *Machine Learning*, estatística inferenciais ou estatísticas descritivas. Esse processo é o que faz da computação algo tão importante dentro da Ciência de Dados uma vez que a máquina consegue tirar conclusões analisando padrões de forma infinitamente mais rápida que o ser humano. (LEE; Udacity, 2017)

6.5 COMUNICAÇÃO

Por fim a comunicação, onde será justificado e transmitido o significado dos *insights* encontrados. Se o resultado final da análise for a construção de um sistema, aqui será compartilhado os resultados por meio de *Dashboards* e gráficos. Os resultados podem ser comunicados de formas variadas: via relatórios, slides, e-mails ou mesmo *Dashboards*. (LEE; Udacity, 2017)

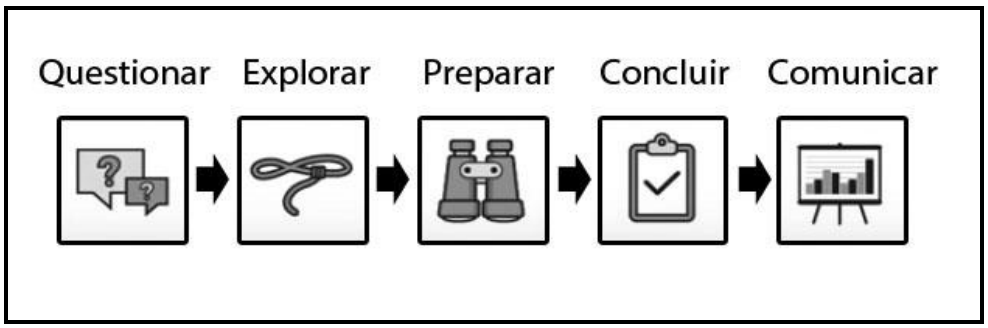

Figura 5: Etapas da Análise de Dados (LEE; Udacity, 2017)

6.6 AMOSTRA DE ANÁLISE DE DADOS

No mundo real existem vários exemplos onde as análises dos dados são realizadas de forma satisfatória, retornando resultados de grande valia para empresas ou ajudando os usuários de um determinado serviço.

A empresa de transmissão de filmes Netflix faz uso da Análise de Dados com Ciência de Dados para recomendar filmes a seus usuários de acordo com seus gostos pessoais.

A empresa de tecnologia Facebook costuma fazer a divulgação de artigos de cunho científico baseados em coleta de dados em sua plataforma, como por exemplo o artigo relacionado a ideologia política dos posts publicados pelas pessoas em seus perfis (BAKSHY; MESSING; ADAMIE, 2017). Até mesmo nos esportes, muitas equipes - independente da modalidade esportiva - fazem o uso de *Data Science* para melhorar o desempenho de seus atletas. Um fato real observado na esporte parte da história transformada em filme de um estatístico americano que fez uso da Análise de Dados para conduzir uma equipe acostumada a derrotas a seguir um caminho de vitórias. (BENNET, 2011).

7 FERRAMENTAS

Para o funcionamento eficaz da aplicação foi utilizada uma das melhores ferramentas disponíveis hoje para o trabalho com Ciência de Dados; a ferramenta Anaconda.

Anaconda possui um conjunto de diversos pacotes indispensáveis para a união de técnicas diversas, tais como Análise de Dados, programação, *ML* e Mineração de Dados.

Todas as ferramentas utilizadas para conclusão deste projeto são *Open Source* e podem ser facilmente encontradas nos sites oficiais das mesmas. Os pacotes abaixo mencionados contam com uma vasta documentação para interessados em conhecer mais a fundo as ferramentas citadas.

7.1 LINGUAGEM DE PROGRAMAÇÃO

A escolha de Python como linguagem de programação utilizada neste projeto, dar-se-á pelo fato de que a mesma se encontra na categoria de linguagem de propósito geral. Sendo assim é classificada como uma linguagem que “pode ser utilizada tanto para desenvolvimento de programas comerciais quanto de programas científicos” (MANZANO, 2011, p. 19). Sua aplicação é extremamente variada e vai desde um simples desenvolvimento *Web* às implementações com *Data Science*.

Por conta de sua “sintaxe clara e concisa que favorece a legibilidade do código-fonte, tornando a linguagem mais produtiva” (BORGES, 2014, p. 14), Python é uma das linguagens mais bem aceitas no meio científico em geral, principalmente por cientistas de áreas que não estão necessariamente ligadas a Tecnologia da Informação, mas que tenham a necessidade de aplicar a computação em algum momento de suas pesquisas.

Figura 6: Logomarca da linguagem de programação Python (PYTHON, 2018)

Outra característica importante de Python e talvez a principal que a torne tão interessante na produção deste documento é a sua grande coleção de recursos desenvolvidos para fins específicos de uma dada ciência, neste caso recursos voltados para Ciência de Dados.

Python também atende ao conceito de portabilidade, funcionando em arquiteturas variadas.

Resumidamente podemos dizer que:

NILO NEY COUTINHO MENEZES (2014, p. 3)

> Python é uma linguagem de programação poderosa e fácil de aprender, ela possui estrutura de dados de alto nível e uma simples, mas eficiente, abordagem de programação orientada a objetos. Sua elegante sintaxe e tipagem dinâmica, juntamente com seu interpretador nativo, fazem dela a linguagem ideal para *scripting* e o desenvolvimento rápido de aplicações em diversas áreas sob várias plataformas.

7.2 PACOTES PYTHON PARA ANÁLISE DE DADOS

Para se trabalhar com *Data Science* apenas a implementação com a linguagem de programação Python não é o suficiente para solucionar diversos tipos de problemas; para isso é necessário o uso de técnicas adicionais contidas em pacotes que devem ser adicionados ao projeto.

Estes pacotes melhoram a aplicação no que diz respeito a velocidade de processamento, simplicidade de codificação e utilização de recursos específicos de Análise de Dados ou Aprendizagem de Máquina.

Devido sua importância, os diversos pacotes imprescindíveis para uma análise bem-sucedida serão brevemente introduzidos nos subtópicos abaixo.

7.2.1 IPython

O IPython é um terminal interativo similar ao terminal Python Shell comum com o adicional de funcionalidades específicas de destaque de sintaxe e auto incremento de instruções próprias da linguagem Python.

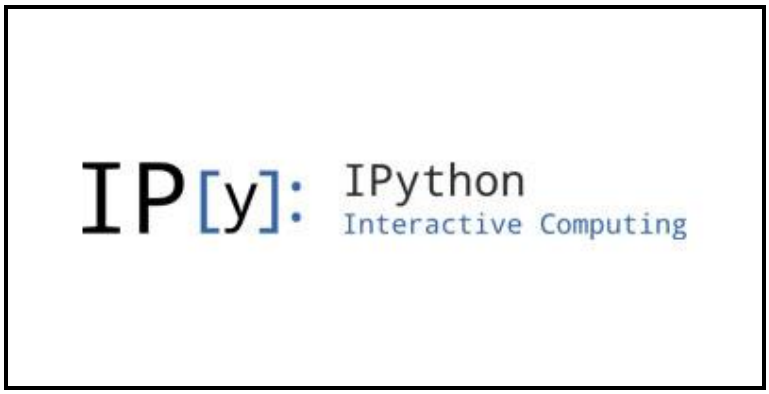

Figura 7: Logomarca do IPython (IPYTHON, 2018)

É geralmente utilizado como núcleo em parceria com Jupyter Notebook já que com IPython todos os arquivos são visualizados em janelas separadas, assim como os vários scripts, funções e classes. (MCKINNEY, 2012, p. 5)

7.2.2 Jupyter Notebook

Jupyter Notebook ou caderno Jupyter é uma ferramenta criada para se trabalhar com programação literária. Neste paradigma de programação há uma intersecção entre a codificação e a documentação em forma de narrativa, ao invés de manipulá-los como elementos independentes.

Figura 8: Logomarca do Jupyter Notebook (NOTEBOOK, 2018)

“Em vez de pensar que nossa tarefa principal é dizer ao computador o que fazer, vamos nos concentrar em explicar aos outros seres humanos o que queremos que o computador faça.” (KNUTH, 1999).

Criado com o objetivo de proporcionar a maior legibilidade humana possível funciona enviando mensagens do navegador *Web* (*browser*) para o núcleo IPython (que roda em segundo plano). O núcleo executa o código e devolve ao notebook que o salva como um arquivo *JSON* no servidor do notebook, podendo ser acessado a qualquer momento.

A ilustração a seguir demonstra de forma mais clara seu processo de funcionamento dentro da aplicação.

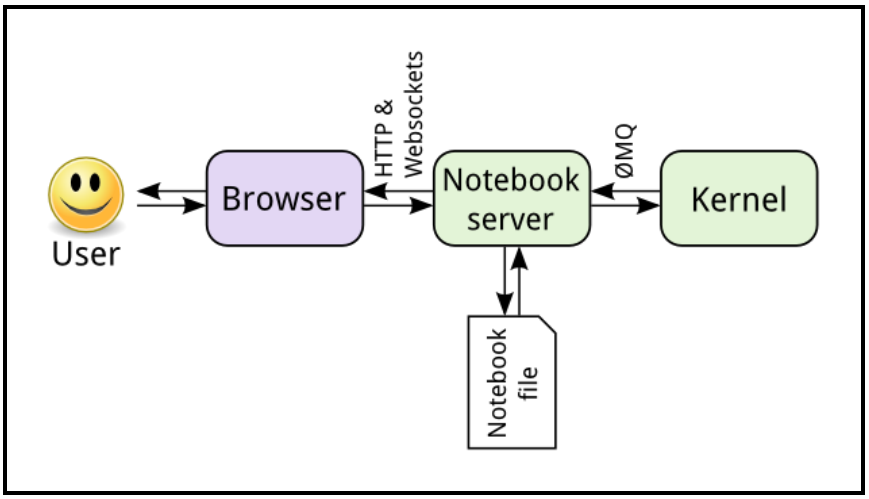

Figura 9: Estrutura Jupyter Notebook

O armazenamento de textos (*HTML*), códigos, imagens e formalismos matemáticos (LaTex) são feitos em um único documento compartilhável.

Seu nome é um acrônimo derivado do nome das linguagens de programação Julia, Python e R. Embora qualquer linguagem possa ser mandada entre o núcleo e o notebook por conta de serem separados, ele é substancialmente mais utilizado com as linguagens Python e R, haja vista que se trata de uma ferramenta essencial para experimentos em *Data Science*. O notebook frequentemente será usado para limpeza de dados, exploração de dados, análise de Big Data, visualizações de histogramas e *Machine Learning*.

7.2.3 Pandas

Fornece uma estrutura de dados e funções avançadas projetada para o trabalho com grandes quantidades de dados de forma mais rápida.

Pandas promove uma rica estrutura de dados e funções desenhadas para tornar mais fácil e rápido o trabalho com um grande conjunto de dados. (MCKINNEY, 2012, p. 4)

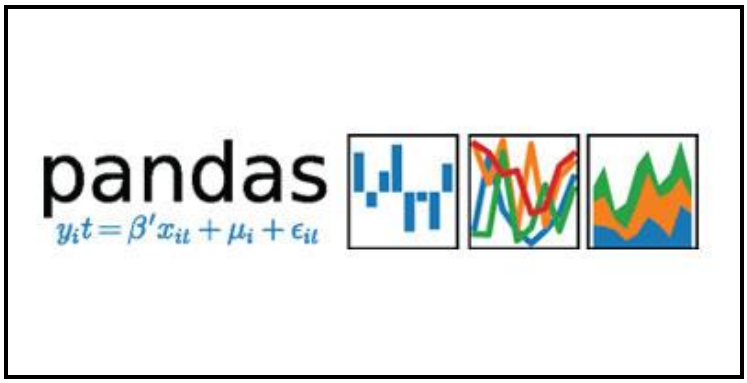

Figura 10: Logomarca do Pandas (PANDAS, 2018)

7.2.4 NumPy

Seu nome se refere a *Numerical Python* e permite calcular com eficiência funções matemáticas de grandes *arrays* (multidimensionais). O NumPy um pacote básico de computação científica.

Assim como a maioria dos pacotes de Python para o uso em Ciência de Dados o Numpy visa facilitar o trabalho do cientista, assim como agilizar os processamentos dos dados em análise. (MCKINNEY, 2012, p. 4)

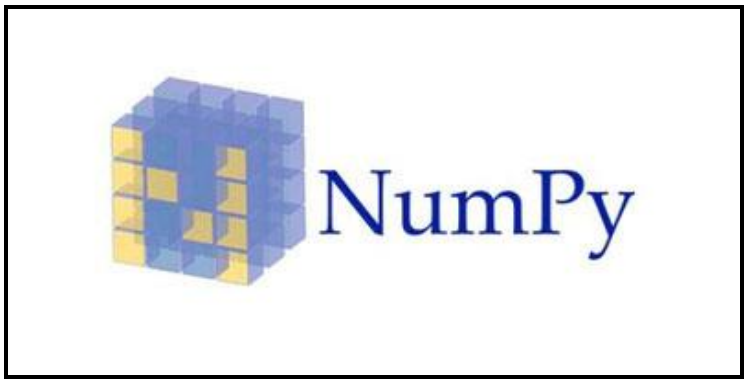

Figura 11: Logomarca do NumPy (NUMPY, 2018)

7.2.5 Matplotlib

Biblioteca popular para a produção de gráficos e outras visualizações 2D. Matplotlib se integra muito bem com Python proporcionando assim um ambiente interativo confortável para a exploração de dados em Python. Esta ferramenta se integra muito bem com o IPython e fornece um ambiente interativo confortável para a exploração de dados e geração de gráficos. (MCKINNEY, 2012, p. 5)

Figura 12: Logomarca do Matplotlib (MATPLOTLIB, 2018)

7.2.6 Scikit-Learn

Para o uso de *ML* na busca de uma análise preditiva tendo como base um banco de dados com um volume substancialmente grande (*Big Data*), o a adição de uma ferramenta de auxílio é de extrema importância, já que se trata de uma forma de codificação que ultrapassa a trivialidade.

Uma das ferramentas para este fim de melhor aceitação no mercado de *software* sem dúvidas é o Sciki-Learn.

O Scikit-Learn é uma ferramenta simples para executar o *Data Mining* com o *Data Analysis* em uma aplicação.

Figura 13: Logomarca do Scikit-Learn (SCIKIT-LEARN, 2018)

7.2.7 Anaconda

Todas essas ferramentas podem ser instaladas de uma só vez através do Anaconda.

O Anaconda é uma distribuição que fornece diversos pacotes que podem ser instalados de uma só vez e que são de extrema importância para se trabalhar com *Data Science*.
Além de já instalar os pacotes, livrando o desenvolvedor de um árduo e demorado trabalho de organização e configuração de ambiente o Anaconda também disponibiliza o Conda, que é responsável pelo controle das versões dos pacotes instalados; permitindo assim ao desenvolvedor trabalhar em diferentes projetos com diferentes versões de Python por exemplo.

Figura 14: Logomarca do Anaconda (ANACONDA, 2018)

8 ESTUDO DE CASO

O foco central deste projeto é o de desenvolver uma Análise de Dados baseado em um conjunto de dados real, disponibilizado por um *Data Set* de livre acesso para estudos.
Dentro da análise encontrar-se-á características relevantes sobre esse conjunto de dados que possam ser exploradas e devolvidas ao analisador em forma de *Insights.*

8.1 VISÃO GERAL DO CASO

Neste exemplo foi selecionado um caso simples, de modelo estruturado e com volume quantitativo substancialmente pequeno para facilitar a visualização das informações.
O modelo aqui estudado refere-se a um conjunto de dados relacionado ao câncer de mama, onde tumores foram avaliados e diagnosticados como malignos ou benignos.
O *Data Set* ou conjunto de dados acima referido é livre para estudos e disponibilizado pela Universidade de Wisconsin (WOLBERG, STREET, MANGASARIAN, 1995).

8.2 ANÁLISE DO CASO

O fluxo completo da análise dos dados, como já foi mencionado no capítulo 6 deste trabalho, é composto por 5 etapas que vão do questionamento a comunicação. Abaixo serão especificados cada um deles usando como referência os diagnósticos dos pacientes do *Data Set* coletado.

8.1.1 Iniciando a análise

Ao iniciar o projeto a primeira ação a se tomar é a de definir como será estruturado o *Data Frame*, ou seja, se a análise partirá de um conjunto de dados existente, onde haja uma estrutura de banco pré-definida; ou se a estrutura do banco será moldada a partir do questionamento, o que tornaria os dados mais preparados desde o início para um uso mais refinado de *Data Science*.
Claramente podemos perceber que a tendência é a de que na maioria das vezes o trabalho será iniciado a partir de um banco de dados já existente, o que levaria o cientista a buscar os campos mais relevantes no que ele considera mais importante no âmbito da área de pesquisa em questão.

Como trabalhar com dados já existentes é a forma mais recorrente, este será o método utilizado no projeto.

8.1.2 Questionar

A partir de um conjunto de dados existente, a tarefa do cientista a principio será a de abstrair informações relevantes partindo de uma observação do mesmo.

Na tabela seguinte pode-se observar que os atributos da tabela guardam informações a respeito de medidas extraídas de diversas mamografias.

	id	diagnosis	texture_mean	perimeter_mean	area_mean	Smoothness_mean	compactness_mean	...
0	842302	M	NaN	122.80	1001.0	0.11840	0.27760	...
1	842517	M	17.77	132.90	1326.0	0.08474	0.07864	...
2	84300903	M	21.25	130.00	1203.0	0.10960	0.15990	...
3	84348301	M	20.38	77.58	386.1	NaN	0.28390	...
4	84358402	M	14.34	135.10	1297.0	0.10030	0.13280	...
5	843786	M	15.70	82.57	477.1	0.12780	0.17000	...
6	844359	M	19.98	119.60	1040.0	0.09463	0.10900	...
7	84458202	M	20.83	90.20	577.9	0.11890	0.16450	...
8	844981	M	21.82	87.50	519.8	0.12730	0.19320	...
9	84501001	M	24.04	83.97	475.9	0.11860	0.23960	...

Tabela 1: Visualização inicial do *Data Frame*

Neste primeiro momento com base na identificação dos pacientes, no tipo de diagnóstico resultante (M para maligno ou B para benigno) e nas medidas coletadas, é importante formular uma pergunta interessante que deva ser explorada. Neste caso podemos questionar se:

- As medidas do tumor podem determinar a magnitude do mesmo?

E em um segundo momento, aprofundando mais o questionamento com:

- Qual ou quais medidas devem ser escolhidas como as principais?

Para não aprofundar o nível de complexidade do projeto, estes serão os únicos questionamentos abordados.

Após avaliar metodicamente os dados disponíveis é possível responder que:

- Sim, as medidas podem determinar a magnitude do tumor, pois existe um certo padrão encontrado entre os tumores malignos e entre os benignos.
- A medida escolhida como a principal será a área média do tumor, pois esta é resultado final entre a medida máxima e mínima da área do tumor encontrada após a doença ser diagnosticada.

Sabendo-se o que queremos encontrar pós-questionamento, é importante melhorar a qualidade do *Data frame* explorando-o e o preparando para uma futura análise mais aprofundada.

8.1.3 Preparar

A preparação é subdividida em 3 novas etapas que são: coletar, avaliar e limpar.

Coletar: A primeira etapa na aplicação do projeto de análise com *Data Science* após conhecer e questionar, ocorre com a coleta. Aqui os dados podem ser adquiridos de formas variadas, tais como: download direto, envio de uma *API*, busca em página *Web*, acesso a um Banco de Dados, entre outros.

Neste caso os dados foram baixados de um *Data Set* disponível para estudos, no formato *CSV*. O Pandas é usado por possuir suporte a grandes quantidades de dados quando necessário.

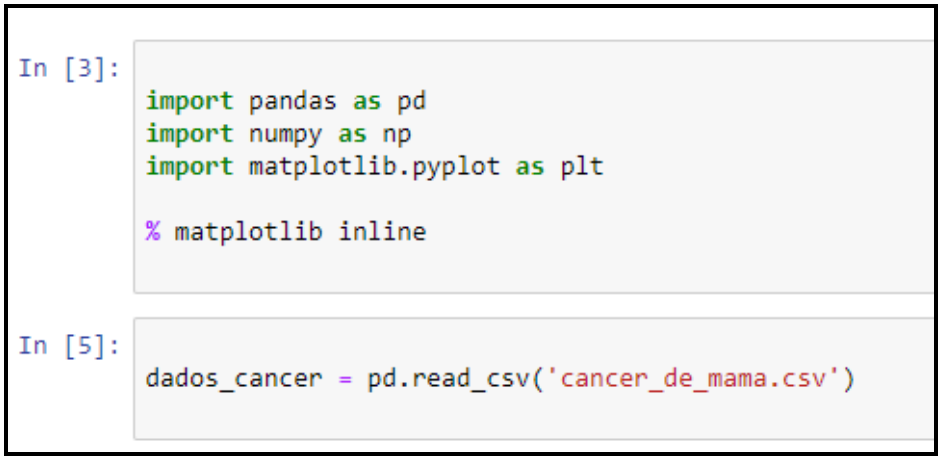

Figura 15: Adicionando os dados

Como podemos ver na visualização dos dados (Tabela 1) por conta de o *Data Set* ser oriundo de uma universidade estadunidense, as indicações dos campos se encontram em inglês. Apenas como forma de melhor entendimento é interessante mudar a língua padrão do *Data Set* sento assim o aportuguesamento dos campos também faz parte da coleta.

```
In [5]:
dados_cancer.rename(columns = {'diagnosis' : 'diagnostico',
                               'radius_mean' : 'raio_medio',
                               'texture_mean' : 'textura_media',
                               'perimeter_mean' : 'perimetro_medio',
                               'area_mean' : 'area_media',
                               'smoothness_mean' : 'suavidade_media',
                               'compactness_mean' : 'compacidade_media',
                               'concavity_mean' : 'concavidade_media',
                               'concave_points_mean' : 'pontos_concavos_medios',
                               'symmetry_mean' : 'simetria_media',
                               'fractal_dimension_mean' : 'dimensao_fractual_media',
                               'radius_SE' : 'raio_SE',
                               'texture_SE' : 'textura_SE',
                               'perimeter_SE' : 'perimetro_SE',
                               'area_SE' : 'area_SE',
                               'smoothness_SE' : 'suavidade_SE',
                               'compactness_SE' : 'compacidade_SE',
                               'concavity_SE' : 'concavidade_SE',
                               'concave_points_SE' : 'pontos_concavos_SE',
                               'symmetry_SE' : 'simetria_SE',
                               'fractal_dimension_SE' : 'dimensao_fractual_SE',
                               'radius_max' : 'raio_maximo',
                               'texture_max' : 'textura_maxima',
                               'perimeter_max' : 'perimetro_maximo',
                               'area_max' : 'area_maxima',
                               'smoothness_max' : 'suavidade_maxima',
                               'compactness_max' : 'compacidade_maxima',
                               'concavity_max' : 'concavidade_maxima',
                               'concave_points_max' : 'pontos_concavos_maximos',
                               'symmetry_max' : 'simetria_maxima',
                               'fractal_dimension_max' : 'dimensao_fractual_maxima'
                              }, inplace=True)
```

Figura 16: Aportuguesando os dados

É interessante pensar neste estágio de aportuguesamento, pois muitos projetos podem ser criados usando como base alguma *API* externa ou um *Data Set* estrangeiro.

Avaliar: Na avaliação o foco está na identificação de dados que não são interessantes para uso. A missão do analisador neste momento é o de apenas encontrar inconsistências, sobretudo acerca de dados faltantes ou dados duplicados.

A Figura 17 traz a instrução necessária para encontrar dados faltantes (*null*); no caso como o *Data Frame* possui 569 entradas (de 0 a 568) todos que obtiverem menos de 569 *non-null* terão dados nulos.

```
In [8]: dados_cancer.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
id                          569 non-null int64
diagnostico                 569 non-null object
raio_medio                  569 non-null float64
textura_media               548 non-null float64
perimetro_medio             569 non-null float64
area_media                  569 non-null float64
suavidade_media             521 non-null float64
compacidade_media           569 non-null float64
concavidade_media           569 non-null float64
pontos_concavos_medios      569 non-null float64
simetria_media              504 non-null float64
dimensao_fractual_media     569 non-null float64
raio_SE                     569 non-null float64
textura_SE                  548 non-null float64
perimetro_SE                569 non-null float64
area_SE                     569 non-null float64
suavidade_SE                521 non-null float64
compacidade_SE              569 non-null float64
concavidade_SE              569 non-null float64
pontos_concavos_SE          569 non-null float64
simetria_SE                 504 non-null float64
dimensao_fractual_SE        569 non-null float64
raio_maximo                 569 non-null float64
textura_maxima              548 non-null float64
perimetro_maximo            569 non-null float64
area_maxima                 569 non-null float64
suavidade_maxima            521 non-null float64
compacidade_maxima          569 non-null float64
concavidade_maxima          569 non-null float64
pontos_concavos_maximos     569 non-null float64
simetria_maxima             504 non-null float64
```

Figura 17: Dados nulos

No exemplo acima os campos textura média, suavidade média, simetria média, textura SE, suavidade SE, simetria SE, textura máxima, suavidade máxima e simetria máxima possuem dados nulos.

Os dados duplicados são transmitidos pelo comando abaixo informado e traz quantos e quais são estes dados duplicados dentro escopo que se deseja analisar.

```
In [9]: sum(dados_cancer.duplicated())

Out[9]: 5

In [10]: dados_cancer.duplicated().tail(15)

Out[10]: 554    False
         555    False
         556    False
         557    False
         558     True
         559    False
         560    False
         561    False
         562    False
         563    False
         564    False
         565    False
         566    False
         567    False
         568    False
         dtype: bool
```

Figura 18: Dados duplicados

No final da avaliação saber-se-á quais os problemas de qualidade encontrados do conjunto de dados, facilitando assim a etapa seguinte de limpeza do mesmo.

Limpar: A limpeza é o estágio final de preparação dos dados, nela tanto os problemas com dados nulos, como duplicados serão solucionados.
Para eliminar dados nulos o Pandas faz uso do método “.fillna()”, onde é preenchido os dados faltantes (NaN) com uma média relacionada aos dados que estão preenchidos, como no exemplo abaixo.

```
In [14]:
dados_cancer['textura_media'] = dados_cancer['textura_media'].fillna(media)
dados_cancer['suavidade_media'].fillna(dados_cancer['suavidade_media'].mean(), inplace=True)
dados_cancer['simetria_media'].fillna(dados_cancer['simetria_media'].mean(), inplace=True)
dados_cancer['textura_SE'].fillna(dados_cancer['textura_SE'].mean(), inplace=True)
dados_cancer['suavidade_SE'].fillna(dados_cancer['suavidade_SE'].mean(), inplace=True)
dados_cancer['simetria_SE'].fillna(dados_cancer['simetria_SE'].mean(), inplace=True)
dados_cancer['textura_maxima'].fillna(dados_cancer['textura_maxima'].mean(), inplace=True)
dados_cancer['suavidade_maxima'].fillna(dados_cancer['suavidade_maxima'].mean(), inplace=True)
dados_cancer['simetria_maxima'].fillna(dados_cancer['simetria_maxima'].mean(), inplace=True)
```

Figura 19: Preenchimento de dados nulos

Já os duplicados são avaliados pelo Pandas e caso alguma linha seja idêntica a outra, o comando ".drop_duplicates()" elimina automaticamente todas as linhas duplicadas. Exemplo:

```
In [16]:
dados_cancer.drop_duplicates(inplace=True)
```

Figura 20: Eliminar de dados duplicados

O resultado das 3 etapas de preparo é visível mesmo quando se observa uma quantidade pequena de dados. No caso deste projeto a partir do preparo os campos passaram a ter nomes em português, dados que antes eram nulos (NaN) como por exemplo texture_mean do paciente 84302 e smoothness_mean do paciente 84348301, agora possuem dados preenchidos com a média geral. Os dados idênticos que são classificados como duplicados também foram removidos por não se fazerem mais necessários a partir deste momento da análise.

A tabela 2 encontrada abaixo mostra de forma clara um conjunto de dados bem moldado e pronto para ser explorado pelo analista de dados, vale lembrar que o não preparo dos dados deve acarretar em perda de precisão a partir do processo de exploração de dados, o que não é interessante na busca de uma Análise de Dados precisa.

	id	diagnostico	texture_media	perimetro_medio	area_media	suavidade_media	compacidade_media	...
0	842302	M	19.293431	122.80	1001.0	0.118400	0.27760	...
1	842517	M	17.770000	132.90	1326.0	0.084740	0.07864	...
2	84300903	M	21.250000	130.00	1203.0	0.109600	0.15990	...
3	84348301	M	20.380000	77.58	386.1	0.096087	0.28390	...
4	84358402	M	14.340000	135.10	1297.0	0.100300	0.13280	...
5	843786	M	15.700000	82.57	477.1	0.127800	0.17000	...
6	844359	M	19.980000	119.60	1040.0	0.094630	0.10900	...
7	84458202	M	20.830000	90.20	577.9	0.118900	0.16450	...
8	844981	M	21.820000	87.50	519.8	0.127300	0.19320	...
9	84501001	M	24.040000	83.97	475.9	0.118600	0.23960	...

Tabela 2: Visualização final do *Data Set*

8.1.4 Explorar

A análise exploratória dos dados é onde começa a observação dos dados na forma de informação. A partir desse estágio o analisador passa a unir conhecimento próprio individual afim de localizar padrões ou informações relevantes se guiando pelo *Data Frame* já preparado para tal.

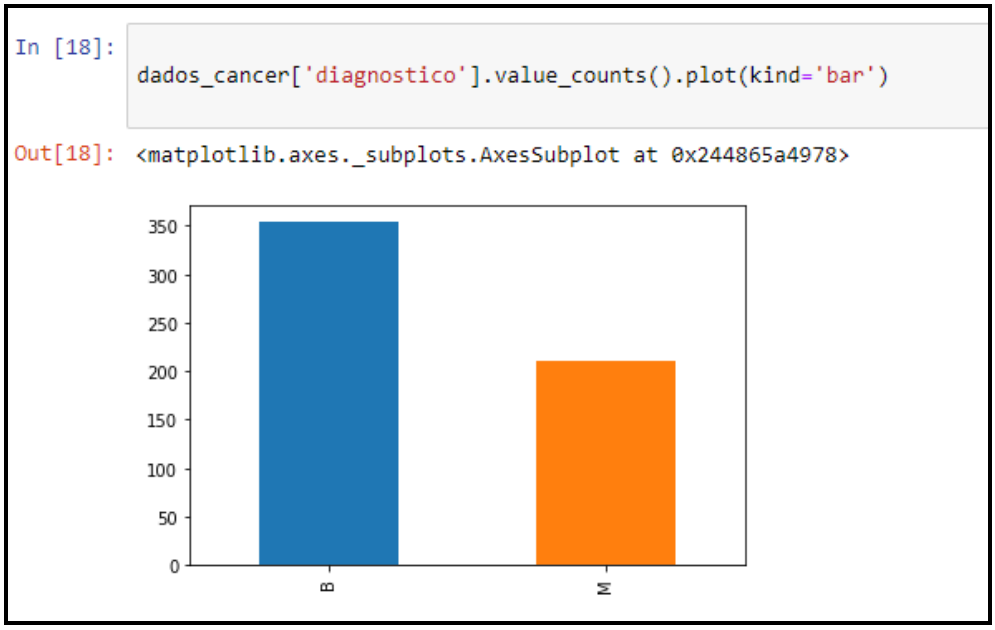

Figura 21: Tumores maligno x benigno

O pacote Pandas possui o comando “.describe()” que deve ser utilizado para a geração de estatística descritiva, como no exemplo abaixo de estatística descritiva para tumores malignos e benignos.

```
In [20]:
dados_cancer_benigno = dados_cancer[dados_cancer['diagnostico'] == 'B']
dados_cancer_benigno['area_media'].describe()

Out[20]: count    354.000000
         mean     462.712429
         std      134.769158
         min      143.500000
         25%      374.975000
         50%      458.150000
         75%      551.550000
         max      992.100000
         Name: area_media, dtype: float64

In [21]:
dados_cancer_maligno = dados_cancer[dados_cancer['diagnostico'] == 'M']
dados_cancer_maligno['area_media'].describe()

Out[21]: count     210.000000
         mean      976.582857
         std       365.494289
         min       361.600000
         25%       706.850000
         50%       932.000000
         75%      1200.750000
         max      2501.000000
         Name: area_media, dtype: float64
```

Figura 22: Estatística descritiva

A estatística descritiva auxilia na tomada de decisão em relação ao tipo de dado estatístico que tomaremos como base.

No retorno do método “describe”, um bom modelo de discrepância entre os tipos de diagnósticos maligno e benigno se encontra em mean (média), logo este será tomado como unidade de medida para explanar melhor uma visão a partir de um histograma unindo os dois tipos de diagnósticos.

O histograma fornecido pelo Matplotlib torna mais visível a relação encontrada entre os altos valores médios dos tumores e seus diagnósticos como malignos ou benignos.

```
In [31]:
fig, ax = plt.subplots(figsize=(15, 10))
ax.hist(dados_cancer_benigno['area_media'], alpha=0.5, label='BENIGNO')
ax.hist(dados_cancer_maligno['area_media'], alpha=0.5, label='MALIGNO')
ax.set_title('Distribuição de Benigno & Maligno pela area do tumor')
ax.set_xlabel('ÁREA MÉDIA')
ax.set_ylabel('QUANTIDADE')
ax.legend(loc='upper right')
plt.show()
```

Figura 23: Impressão de histogramas

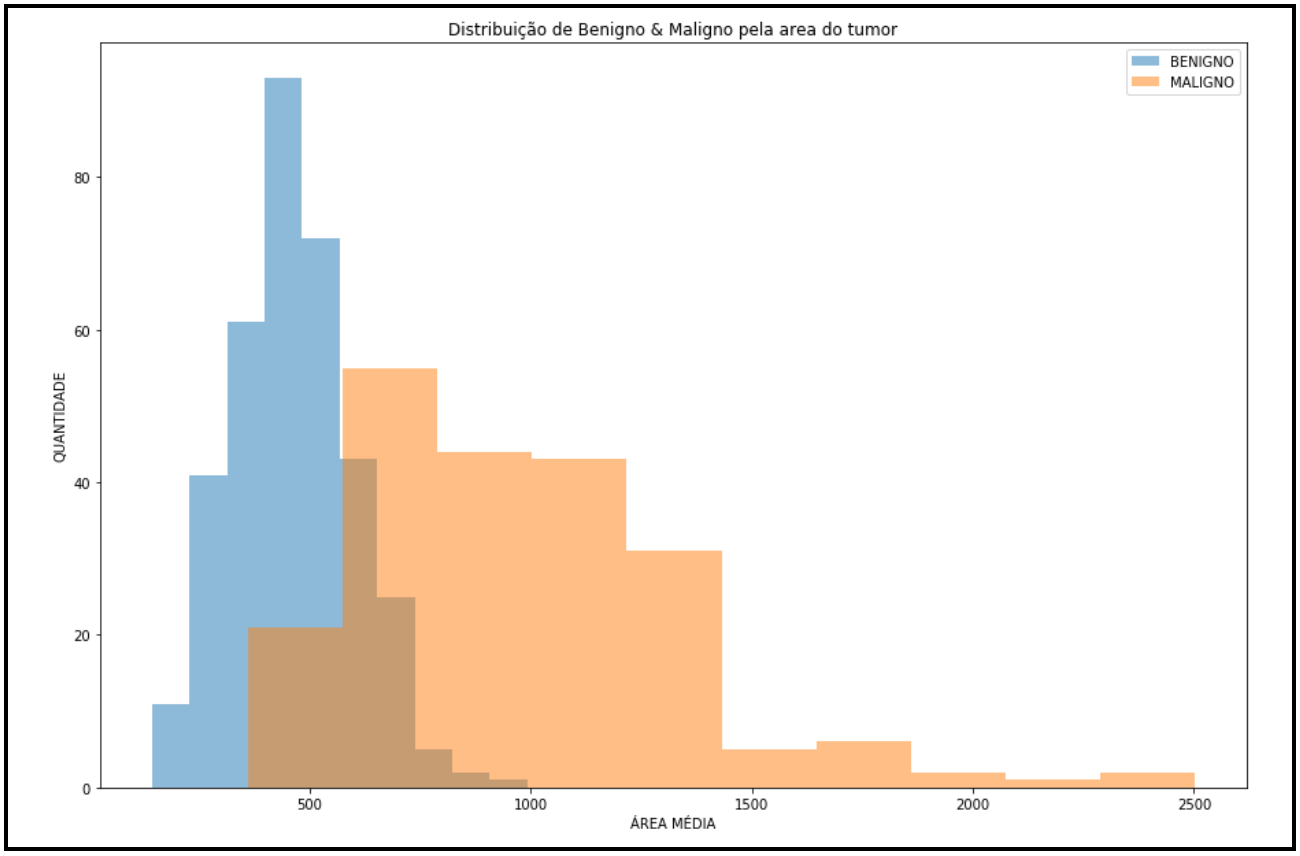

Figura 24: Histogramas maligno x benigno

8.1.5 Concluir

A conclusão tem como característica principal o uso de recursos computacionais que potencializam o trabalho do cientista; é justamente neste estágio que se emprega o uso

do pacote Scikit-Learn com o *Machine Learning* no reconhecimento de padrões, buscando abstrair qualquer *Insight* válido e lógico sobre uma determinada análise.

Inúmeras técnicas podem ser utilizadas para a Aprendizagem de Máquina, neste caso a escolhida foi a de K-enésimo vizinho mais próximo ou *K-NN* (*K-Nearest Neighbor*). Seu funcionamento se baseia em identificar um elemento a ser estudado, comparando-o a outros elementos cujas características (dados) sejam de alguma forma semelhantes (padrões).

No exemplo abaixo pode-se observar que os elementos de análise (estrela), são caracterizados com as mesmas propriedades de seus 4 vizinhos mais significativos.

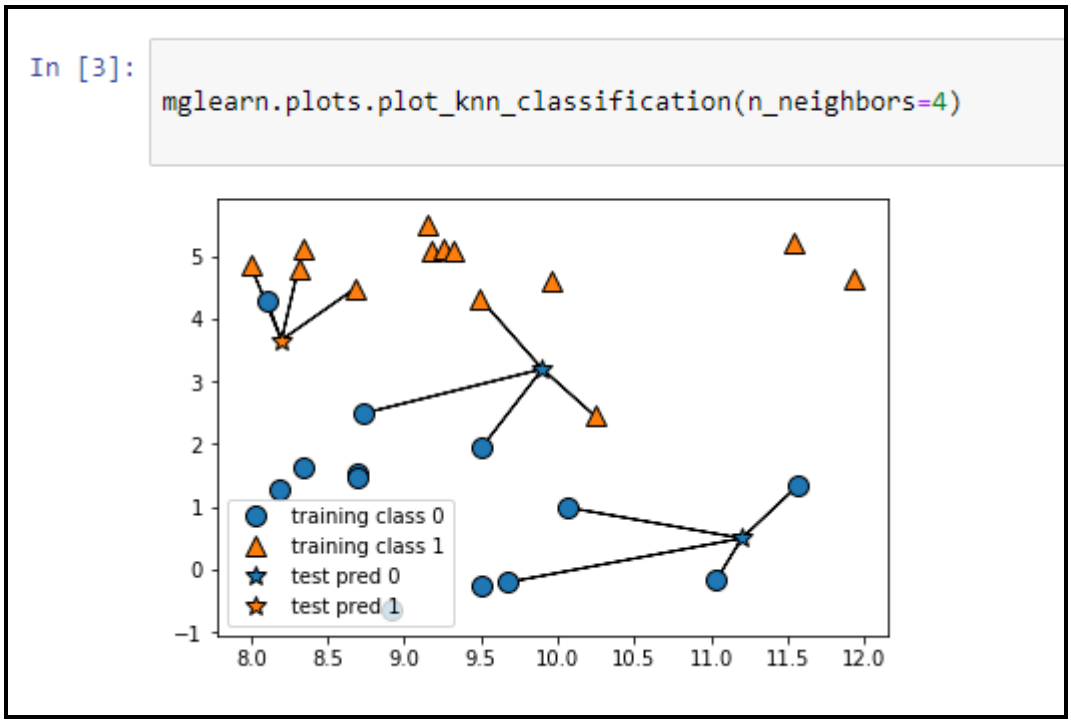

Figura 25: Exemplo de K-NN com 4 vizinhos

No caso deste projeto o *ML* com a técnica de *k-NN,* vai utilizar o conjunto de dados para retornar à capacidade preditiva de um elemento específico quando informado.

Este tipo de implementação vai nos servir para definir a precisão do *Data Frame*, e representar qual a confiança que se pode ter nesta aplicação em específico.

Após o estudo de um elemento teste jogado aleatoriamente no sistema, o retorno codificação é 0.9230769230769231. Este valor significa que o algoritmo possui uma precisão de mais de 92% em uma análise K-NN.

```
In [13]:
from sklearn.neighbors import KNeighborsClassifier

def funcao_quantidade():
    X_train, X_test, y_train, y_test = answer_four()

    resultado = KNeighborsClassifier(n_neighbors = 4)
    resultado.fit(X_train, y_train)
    return resultado
funcao_vizinhos()
```

```
Out[13]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                              metric_params=None, n_jobs=1, n_neighbors=4, p=2,
                              weights='uniform')
```

```
In [17]:
def precisao():
    X_train, X_test, y_train, y_test = answer_four()
    resultado = funcao_quantidade()
    return vizinhos.score(X_test, y_test)
answer_eight()
```

```
Out[17]: 0.9230769230769231
```

Figura 26: Definindo precisão com *K-NN*

8.1.6 Comunicar

A comunicação finaliza todo o processo da Análise de Dados com *Data Science*; nesta etapa algum tipo de conhecimento encontrado deve ser transmitido de forma que seja visível ao receptor da mensagem.

Um bom meio de transmitir essas informações é por intermédio de gráficos fornecidos pelo Matplotlib e textos descritivos.

Na etapa de exploração foi possível abstrair conhecimento importante acerca do faixa de risco primordial de tratamento e a faixa intermediária (Figura 27) que exibe os pacientes que exigem um grau de atenção elevado.

O conjunto de dados utilizados tendem a retornar resultados satisfatórios caso seja utilizado a técnica de *K-NN* informando os dados de um novo paciente; sua porcentagem de acerto ultrapassa os 92%.

Mesmo que pequena o uso de *Data Science,* mostra sua de extrema importância no objetivo de se otimizar a análise.

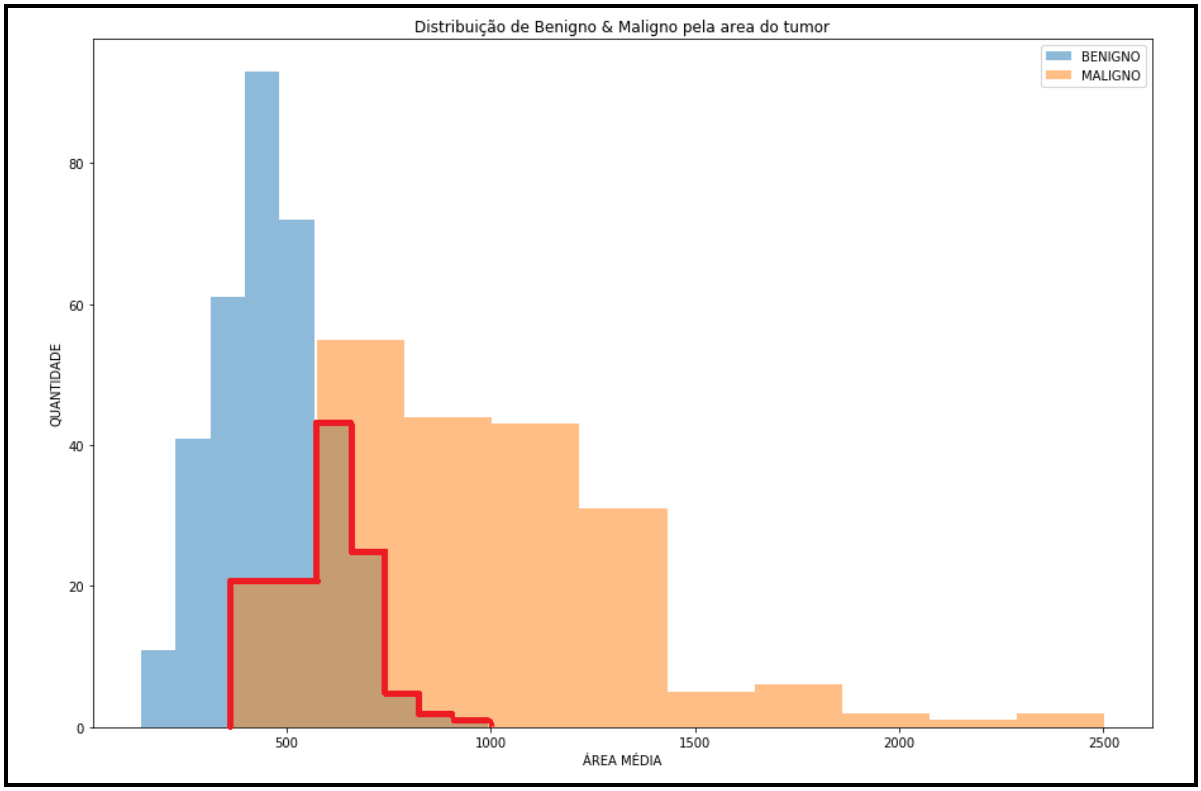

Figura 27: Taxa de risco

9 CONCLUSÃO

9.1 CONSIDERAÇÕES FINAIS

Percebe-se que para a utilização de *Data Science* é necessário possuir uma gama enorme de competências dentro de diversas áreas do conhecimento, tais como: programação, matemática, estatística, Análise de Dados, entre outras. Por conta disto a conclusão de uma aplicação final é algo extremamente demorado e que demanda bastante empenho por conta do cientista.

No escopo deste projeto, todos os temas mencionados tiveram de ser estudados com a finalidade de incrementar o desenvolvimento da aplicação final.

Apesar da necessidade de competências multidisciplinares, é provável deduzir que mesmo com uma aplicação simples, é possível extrair resultados relevantes que possam ser replicados em organizações de portes variados. Esta característica torna viável a simulação de um sistema com *Data Science* em diversos ramos onde a computação se aplica, fato este que valida objetivo inicial deste trabalho.

Em fatos gerais, o uso de *Data Science* pode facilitar muito a vida de organizações que tenham a filosofia de incorporar alta tecnologia, na busca de melhores resultados.

9.2 PROJETOS FUTUROS

Sabendo-se que a capacidade de precisão na predição de um paciente dentro deste sistema é de mais que 92%, uma ideia interessante a partir desta implementação será o uso de um algoritmo que informe a possibilidade de um tumor encontrado em um paciente “x”, ser identificado como maligno ou benigno tendo como base as medidas deste tumor. Para este fim a técnica de *Machine Learning* necessita ser melhor explorada e incorporada no ambiente Jupyter Notebook.

Mesmo com todas as funcionalidades da Ciência de Dados incorporadas no projeto, muito dos recursos utilizados precisam ser melhor observados, o que acarretaria em pesquisas científicas com resultados surpreendentes, ao ponto de proporcionar a sociedade resultados que possam prosperar a vida da comunidade em geral.

REFERÊNCIAS

ANACONDA. **Site Oficial do Pacote Anaconda para *Data Science*.** Disponível em https://anaconda.org/ Acesso em: 20 de maio de 2018

BAKSHY, Eytan; MESSING, Solomon; ADAMIE, Lada A. **Exposure to Ideologically Diverse News and Opinion on Facebook.** Facebook. Disponível em https://research.fb.com/wp-content/uploads/2015/05/exposure-to-ideologically-diverse.pdf?. Acesso em: 05 de março de 2018

BENNETT, Miller. **Moneyball: O Homem que Mudou o Jogo**, Columbia Pictures, 2011

BORGES, Luiz Eduardo. **Python para Desenvolvedores,** 1. ed. São Paulo – SP: Editora Novatec, 2014

BOSCHETTI, Alberto; MASSARON, Luca. **Python Data Science Essentials**, 2. ed. Birmingham – UK: Editora Packt, 2016

BRAGA, Luís Paulo Vieira; **Introdução à Mineração de Dados**, 2. ed. Rio de Janeiro – RJ: Editora E-Papers, 2005

CIELEN, Davy; MEYSMAN, Arno D. B.; ALI, Mohamed. **Introducing Data Science: Big Data, Machine Learning, and more using Python Tools,** 1. pub. Shelter Island - Estados Unidos: Editor Manning Publication, 2016

DATE, Christopher J. **Introdução a Sistemas de Banco de Dados**, 8. ed. Tradução de Daniel Vieira. Rio de Janeiro – RJ: Editora Campus, 2003

ELRASMI, Ramez; NAVATHE, Shamkant B. **Sistema de Banco de Dados,** 6. ed. Tradução de Daniel Vieira. São Paulo – SP: Editora Pearson, 2011

FAWCETT, Tom; PROVOST Foster. **Data Science for Business,** 1. ed., Sebastopol - Estados Unidos da América: Editora O'Reilly, 2013

IPYTHON. **Site Oficial do Pacote IPython.** Disponível em https://ipython.org/ Acesso em: 19 de maio de 2018

KNOTH, Donald. **Literate Programming**. Disponível em www.literateprogramming.com. 1992, pg. 99, Acesso em: 16 de junho de 2018

LEE, Juno; Udacity. **Análise de Dados.** Disponível em https://www.youtube.com/watch?v=qdV4sifMmWI. Acesso em: 23 de maio de 2018.

MANZANO, José Augusto N. G. **Programação de Computadores com C++,** 1. ed. 3. Reimpressão. São Paulo – SP: Editora Érica, 2011

MARQUESONE, Rosângela. **Big Data: Técnicas e Tecnologias para Extração de Valores dos Dados,** 1. ed. São Paulo – SP: Editora Casa do Código, 2016

MATPLOTLIB. **Site Oficial do Pacote Matplotlib.** Disponível em https://matplotlib.org/gallery/api/logos2.html Acesso em: 21 de maio de 2018.

MCKINNEY, Wes. **Python for Data Analysis**, 1. ed. Sebastopol – CA: Editora O'Reilly, 2012

MENEZES, Nilo Ney Coutinho. **Introdução à Programação com Python,** 2. ed. 3. Reimpressão. São Paulo – SP: Editora Érica, 2014

NOTEBOOK; Jupyter. **Site Oficial da IDE de Jupyter Notebook.** Disponível em http://jupyter.org/ Acesso em: 19 de maio de 2018

NUMPY. **Site Oficial do Pacote NumPy.** Disponível em http://www.numpy.org/ Acesso em: 20 de maio de 2018

OLAVSRUD, Thor. **Afinal o que é Ciência de Dados e o que isso tem a ver com a profissão do futuro.** Disponível em http://idgnow.com.br/carreira/2018/07/05/afinal-o-que-e-ciencia-de-dados-e-o-que-isso-tem-a-ver-com-as-profissoes-do-futuro/ Acesso em: 10 de janeiro de 2018

PANDAS. **Site Oficial do Pacote Pandas.** Disponível em https://pandas.pydata.org/ Acesso em: 20 de maio de 2018

PIERSON, Lillian. **Data Science for Dummies,** 2. ed. Hoboken - Estados Unidos da América: Editora Wiley, 2017

PYTHON. **Site Oficial da Linguagem de Programação Python.** Disponível em https://www.python.org/community/logos/ Acesso em: 18 de maio de 2018.

RICHERT, Willi; COELHO Luis Pedro. **Building Machine Learning Systems with Python,** 2 ed. Birmingham – UK, Editora Packt, 2013

SCIKIT-LEARN. **Site Oficial do Pacote Scikit-Learn.** Disponível em http://scikit-learn.org/stable/ Acesso em: 22 de maio de 2018.

SETZER, Waldemar W. **Dado, Informação, Conhecimento e Competência**, Revista de Ciência da Informação, n.0, dezembro, 1999, artigo 1 https://www.ime.usp.br/~vwsetzer/datagrama.html

THIERAUF, Robert J. **Knowledge Management Systems for Business,** 1. ed. Londres – Reino Unido: Editora Quorum Books, 1999

VANDERPLAS, Jake. **Python Data Science Handbook,** 1. ed. Sebastopol – Estados Unidos: Editora O'Reilly Media, 2016

WEIS, Sholom M.; INDURKHYA Nitim. **Predict Data Mining**, 1. ed. Editora Morgan Kaufmann Publishers Inc, 1999

WILLIAMSON, Jason. **Getting a Big Data Job for Dummies,** 2. ed. Hoboken - Estados Unidos da América: Editora Wiley, 2015.

WOLBERG, William H.; STREET, W. Nick; MANGASARIAN, Olvi L. **Cancer Data Set**, University os Wisconsin – Estados Unidos da América. Disponível em: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). Acessado em: 20 de Fevereiro de 2018