

**PROJETO DE PESQUISA**  
**Laboratório de Ciências Matemáticas (LCMAT - Computação)**  
**Centro de Ciência e Tecnologia (CCT)**

*Project-driven Data Science: Aprendendo e Mapeando*

Proposta submetida ao Edital PIBi 03/2021 - PIBIC NOTA 10

**Proponente:** *Prof<sup>a</sup>. Dra. Annabell Del Real Tamariz,*  
[annabell\\_pos@pq.uenf.br](mailto:annabell_pos@pq.uenf.br) / [annabell@uenf.br](mailto:annabell@uenf.br)  
Telefone: (22) 998315575

**CAMPOS DOS GOYTACAZES, RJ**

**DEZEMBRO DE 2020**

## **RESUMO**

A nossa proposta de pesquisa tem como finalidade oferecer aos alunos do curso de bacharelado em Ciência da Computação da UENF, a possibilidade de adentrar no mundo científico entrelaçando teoria e prática no campo da Ciência de Dados, com a possibilidade de construir uma estrutura de aprendizado por meio do desenvolvimento de subprojetos em Ciência de Dados, usando dados do cotidiano da sociedade e transformando-os em informações e conhecimentos úteis, tornando possível, entre outras funcionalidade, oferecer modelos que ajudem na tomada de decisão inteligente. Está proposta é considerada um projeto de pesquisa aplicada e de inovação e sua importância radica na sugestão de pensar a ciência de dados não como um novo domínio de conhecimento a ser aprendido, mas como um novo conjunto de habilidades que pode se aplicar em qualquer área de especialização, incluindo o estudo de métodos estatísticos, de visualização de dados, aprendizado de máquina, e ferramentas para armazenamento, depuração e análises de dados. Neste projeto o termo “Project-Driven Data Science” está relacionado ao desenvolvimento e execução de projetos que abordem alguns dos tipos fundamentais de problemas em Ciência de Dados, tais como: Classificação, Regressão, Associação e Clustering. O projeto propõe desenvolver, nos estudantes da Computação na UENF, a capacidade de construir uma estrutura de aprendizado por meio do desenvolvimento de projetos em Ciência de Dados, visando abranger as principais aplicações e ferramentas básicas do arsenal de um cientista de dados.

**Palavras Chaves :** Ciência de dados, aprendizado de máquina, inteligência artificial.

## 2. INTRODUÇÃO E JUSTIFICATIVA

Um dos primeiros conceitos que devem ficar claros no desenvolvimento deste projeto de pesquisa é o relacionado à seguinte questão: que é um dado?; O que se entende por informação?; Que seria então o conhecimento? Seguindo o blog do Zouza<sup>1</sup> publicado em 2020, ele traz a seguinte definição para estes termos, que parecem mas não são nada parecidos.

- **Dados:** Corresponde às informações não tratadas, ou seja, dados correspondem a códigos que constituem a matéria prima da informação. Os dados representam um ou mais significados que isoladamente não podem transmitir uma mensagem ou representar algum conhecimento.
- **Informação:** Dados tratados, dotados de relevância e propósito. Eles tem um significado e são organizados por algum propósito.
- **Conhecimento:** Informações que podem ser utilizadas, ou seja, refere-se à habilidade de criar modelos que descrevem o objeto, as ações a implementar e as decisões a tomar.

Partindo desses conceitos iniciais, pode-se dizer que a Ciência de Dados (CD) é a área que busca organizar e expandir o conjunto de ferramentas, técnicas e fundamentos teóricos utilizados no processo de transformar dados brutos em informação e/ou conhecimento relevante e aplicável. Este processo, segundo (OZDEMIR, 2016), é dividido em cinco passos. Cada um desses passos se traduz em diversas técnicas (histograma, box plotting, etc), ferramentas (python, random forest, algoritmos, linguagens etc) e fundamentos teóricos (estatística, parâmetros, Intervalos de confiança, etc):

1. Fazer uma pergunta interessante;
2. Obter dados (Data Collection)
3. Estruturar e explorar os dados (Data Wrangling and Data Analysis)
4. Modelagem dos dados (Machine Learning)
5. Comunicar e visualizar os resultados para público não especializado. (Data Visualization and Deployment)

Nesse projeto de pesquisa propomos percorrer esse caminho, mais de uma vez, desenvolvendo diferentes projetos, e documentando cada etapa, mapeando os conceitos e

---

<sup>1</sup> <https://blogdozouza.wordpress.com/>

métodos. Disponibilizaremos todo o código produzido bem como referências aos recursos adicionais utilizados.

Propomos uma aprendizagem baseada em projetos porque nos parece primordial para as necessidades do presente e futuro uma aprendizagem exploratória, emancipadora diversa e robusta. O que só pode ocorrer ao experimentar com o objeto de estudo inserido em sua comunidade. Ou seja, explorar projetos interessantes estando a par do que tem sido produzido na academia, aplicado no mercado, e discutido na literatura.

A plataforma Kaggle<sup>2</sup> é parte fundamental desse processo, pois é a principal ferramenta de aprendizado prático em CD (*Data Science*), contendo diversos projetos e competições estruturados com *datasets*, objetivos, fóruns de discussão e principalmente a parte de *notebooks* onde cada um pode submeter um notebook nos moldes do Jupyter<sup>3</sup> com a sua solução para o problema. Dessa forma ao praticar com *feedback* imediato, acesso a outras abordagens do mesmo problema, além das discussões que podem aprofundar os conceitos utilizados temos uma maneira formidável de avançar em ferramental, *insights* e experiência.

## INTELIGÊNCIA ARTIFICIAL

Quando se fala em Inteligência Artificial (IA), quem não é da área de computação pensa em um robô realizando tarefas para nós humanos, mas essa área do conhecimento é muito antiga, data de 1956 quando pela primeira vez apareceu o termo IA e inclui mais do que a criação de robôs, (SILVA, 2004). Nas nossas ações rotineiras do dia-a-dia temos inteligência artificial embutida/inserida nas máquinas, por exemplo, na busca de alguma informação/produto/colega no google, na hora de solicitar um uber para movimentar-nos de um lugar a outro, quando nos é recomendado músicas ou filmes, entre outras atividades que realizamos e não percebemos.

Nesta linha de pesquisa aparecem ao longo do tempo 3 termos interligados, vide Figura 1, que significam diferente mas estão contidos um no outro: Inteligência Artificial (IA), Aprendizado de Máquina (ML - *Machine Learning*) e Aprendizagem Profunda (DL - *Deep Learning*)<sup>4</sup>.

---

<sup>2</sup> <https://www.kaggle.com/>

<sup>3</sup> <https://jupyter.org/>

<sup>4</sup> [A Diferença Entre Inteligência Artificial, Machine Learning e Deep Learning](#)

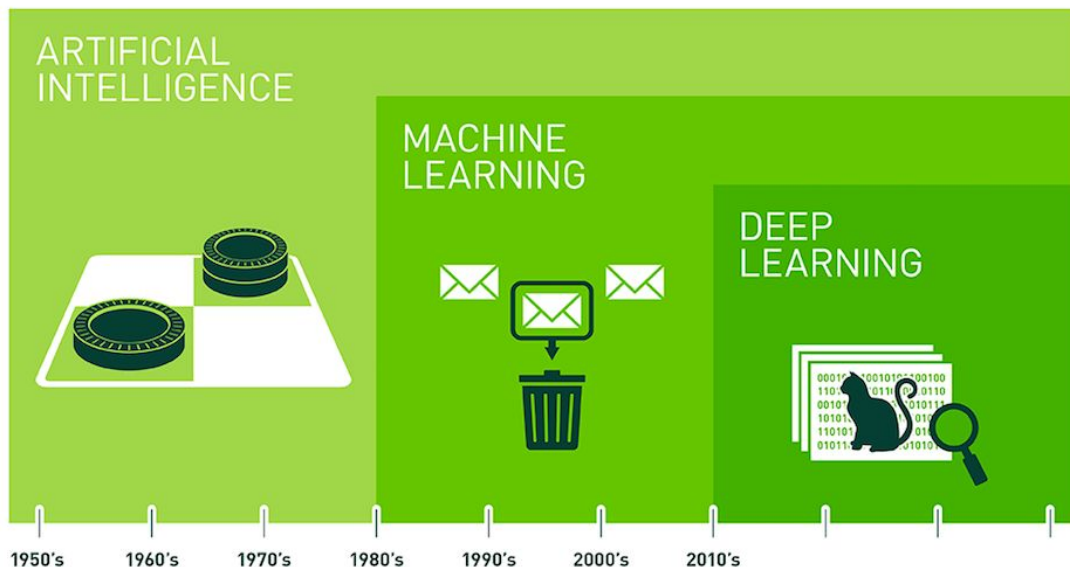


Figura 1: Evolução da IA.

Fonte: [A Diferença Entre Inteligência Artificial, Machine Learning e Deep Learning](#)

Sem entrar muito em detalhes, pode-se definir que IA é o ramo da ciência da computação que estuda como construir máquinas que executam tarefas de forma inteligente, ou seja, além do que foi explicitamente programado. Atualmente, costuma-se denominar IA nas abordagens de problemas com contornos menos precisos, com grande ou infinito número de soluções e possíveis inputs, nos quais busca-se a melhor solução dentro de um universo de resultados possíveis<sup>5</sup>. Já o ML, como o próprio nome sugere, a máquina aprende; usa os algoritmos para coletar informações e/ou dados, aprender com eles e então realizar uma classificação ou predição sobre algum tema em específico do mundo real. Em outras palavras, em lugar dos programadores implementarem programas de software com um conjunto específico de dados, o computador é treinado usando algoritmos com um número significativo de dados, isto cria modelos os quais oferecem à máquina a habilidade de aprender como executar uma determinada tarefa de classificação ou predição, por exemplo. Já a área de Ciência de Dados (CD) foi popularizada a nível empresarial, pois com ajuda dos algoritmos de ML os dados gerados por elas ganharam um valor importante que impulsionam os negócios.

Por esta razão, este projeto de pesquisa aborda a proposta de o estudante escolher que tipo de problema quer resolver. Lembrando que a CD é uma prática em formação, onde devemos nos concentrar no método científico, focando na ciência e não nos dados; usando uma abordagem

<sup>5</sup> <https://towardsdatascience.com/ai-ml-or-data-science-a-glossary-of-terms-529435823bb6>

empírica para obter *insights* a partir dos dados.

## **APRENDIZADO DE MÁQUINAS E/OU CIÊNCIA DE DADOS**

Usando o livro, (VANDERPLAS, 2017) como uma das referências para este projeto, podemos pensar no Aprendizado de Máquina (ML) como um meio que envolve a construção de modelos matemáticos para ajudar a entender os dados, de um determinado problema em análise. O estudo desta linha de pesquisa em IA certamente surgiu de pesquisas e principalmente da aplicação de métodos de aprendizado de máquina da ciência de dados.

O conceito de ML foi descrito em 1959 pelo engenheiro Arthur Samuel (SAMUEL, 1959) e hoje constitui a principal técnica de IA a ser utilizada, graças a dois fatores muito relacionados:

- Crescimento do volume de dados gerados e armazenados, graças a internet.
- Surgimento da Unidade de Processamento Gráfico (GPU) que permite que o processamento das máquinas seja mais rápido e mais barato.

Exemplos de áreas onde pode ser utilizado o tema de [Aprendizado de máquina \(ML\)](#) para resolver problemas específicos (CLAVERA, 2019):

- Sistemas de previsão para empréstimo financeiro; para cálculo de probabilidade.
- Na área de diagnóstico para reconhecer tecidos cancerígenos.
- Na tradução automática de palavras faladas num texto.
- Nas investigações de fraudes e verificações de crédito.

A Ciência da Dados (CD), pode ser definida graficamente usando o diagrama de Venn (SCHUTT;O'NEIL, 2014), como apresentado na Figura 2; e explicitamente como a interseção do conjunto interdisciplinar de habilidades em estatística, computação e negócios, que estão se tornando cada vez mais importantes em muitas aplicações na indústria e na academia. Ou seja, a área da Ciência de Dados é fundamentalmente um assunto interdisciplinar.

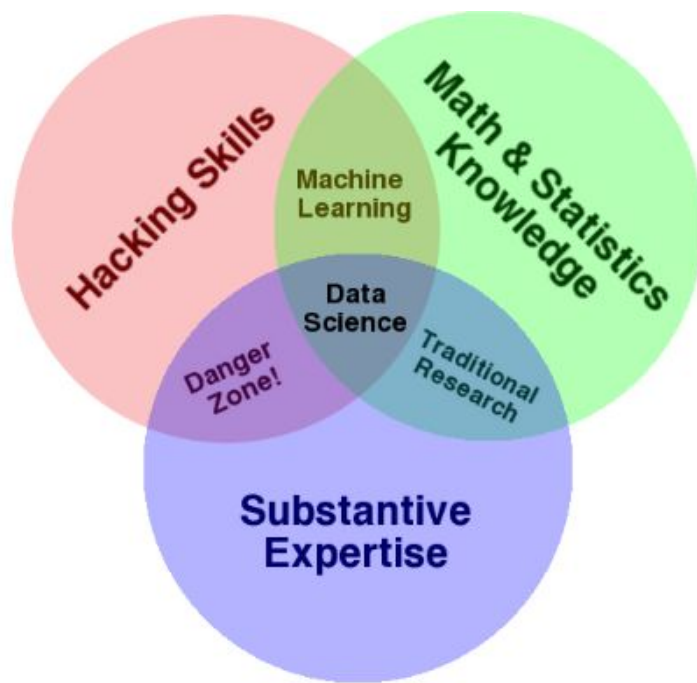


Figura 2: Relação entre áreas para entender a Ciência de Dados. (SCHUTT;O'NEIL, 2014).

A ciência de dados compreende três áreas distintas e sobrepostas: as habilidades de um estatístico que sabe como modelar e resumir conjuntos de dados (que estão cada vez maiores); as habilidades de um cientista da computação que pode projetar e usar algoritmos para armazenar, processar e visualizar esses dados com eficiência; e a experiência de domínio - o que podemos pensar como treinamento “clássico” em um assunto - necessária tanto para formular as perguntas certas quanto para colocar suas respostas no contexto (VANDERPLAS, 2017).

A ciência de dados, nascida da computação e da estatística; é uma prática que atraiu muita atenção, se baseia na transformação de grandes quantidades de dados em conhecimento e informação via previsões e *insights* úteis. Blei e Smyth (2017) discutem no seu artigo a ciência de dados a partir de três perspectivas: estatística, computacional e humana. Embora cada uma das três seja um componente crítico da ciência de dados, os autores acreditam que a combinação eficaz de todos os três componentes é a essência do que trata a ciência de dados. A estatística guia a coleta e análises de dados, já a computação desenvolve algoritmos, onde por exemplo, pode distribuir conjuntos de dados por múltiplos processadores produz equipamentos com grande capacidade de memória, permitindo armazenamento desses dados. Sob a perspectiva humana, a ciência de dados contempla modelos estatísticos e métodos computacionais para resolver problemas específicos de outras disciplinas, entender o domínio

desses problemas, decidir quais dados obter, como processá-los, explorá-los e visualizá-los, selecionar um modelo estatístico e métodos computacionais apropriados, além de comunicar os resultados da análise de uma forma inteligível para aqueles que propuseram os problemas.

Se queremos entender e pensar sobre ciência de dados, devemos pensar em ciência e nunca nos dados; devemos usar os dados para refletir sobre experimentos e questões a serem respondidas sobre esses dados.

Como definido na introdução, a ciência de dados está em definição pela prática o que leva computacionalmente ao uso de softwares. Nesta categoria são apresentadas muitas ferramentas, mas não devemos tentar aprender todas. Tomando como referência a Doug Rose no seu curso “Uso de estatística e software”<sup>6</sup>, existem três categorias de ferramentas que poderiam ser estudadas neste projeto: armazenamento, depuração e análise. Relacionadas ao armazenamento, as mais populares são Hadoop, Cassandra e PostgreSQL. Já na depuração podemos usar ferramentas de script ou linguagens como Python. Finalmente na categoria de análises os pacotes estatísticos como R e SPSS assim como as bibliotecas do Python ajudam bastante.

O objetivo principal deste projeto é fornecer ao estudante bolsista a capacidade de perguntar e responder novas perguntas sobre a área de assunto escolhida, criando uma estrutura de aprendizagem na área da Ciência de Dados. Com a realização deste projeto, teremos um caminho possível sugerido, que entendemos estar de acordo com as maiores necessidades do mundo moderno: uma aprendizagem moderna, eficaz, autodidata, exploradora e interessada. Porque Ciência de Dados é acima de tudo sobre gerar valor, responder perguntas interessantes, abordar problemas complexos, e muitas vezes tentar medir o imensurável. Insistindo sempre que Ciência de Dados é aplicar o método científico aos dados. Visando abranger as principais aplicações e ferramentas básicas do arsenal de um cientista de dados, propomos executar subprojetos que abordem os tipos fundamentais de problemas em Ciência de Dados: Classificação, Regressão, Associação e Clustering.

### **3. MOTIVAÇÃO E OBJETIVOS**

A motivação nesta proposta de pesquisa na área da Ciência de Dados e/ou Aprendizado de Máquina (ML), surge do interesse em encontrar padrões, limpar dados, extrair informação relevante, responder perguntas, estabelecer nexos de causalidade, desenvolver modelos preditivos, encontrar parâmetros relevantes, e até mesmo identificar as limitações de dados,

---

<sup>6</sup> [Uso de estatística e software - Introdução à Ciência de Dados: Conceitos Básicos Tutorial em vídeo](#)



para construir uma base sólida de conhecimento e domínio das principais ferramentas que ajudem na resolução de problemas nas mais diversas áreas da vida.

Para alguns, aplicar ML em grandes volumes de dados, parece ser uma coisa simples e resolve nossos problemas; mas a realidade nas pesquisas demonstram que para os métodos que fazem parte do Aprendizado de Máquina serem eficazes você precisa ter um conhecimento e compreensão dos pontos positivos e negativos de cada um desses métodos assim como dos conceitos gerais que o englobam (GRUS, 2015). Logo este projeto, tem como finalidade oferecer aos alunos do curso de Ciência da Computação, a possibilidade de adentrar neste mundo científico entrelaçando teoria e prática no campo da Ciência de Dados, com a possibilidade de construir uma estrutura de aprendizado por meio do desenvolvimento de projetos de Ciência de Dados, usando projetos e informações na forma de dados do cotidiano da sociedade, tornando possível, entre outras funcionalidade, oferecer modelos que ajudem na tomada de decisão.

Desta forma, acredita-se que este projeto constata sua importância nesta área e como instituição pública, nós pesquisadores e educadores devemos investigar para oferecer propostas à tomada de decisão que os gestores públicos precisam fazer, em especial podemos começar na solução de problemas de interesse social, em relação, por exemplo, a descobrir resultados eleitorais, prevendo retornos de ações, otimizando cliques em anúncios online, identificando microorganismos em fotos de microscópio, buscando novas classes de objetos astronômicos ou trabalhando com dados em qualquer outro campo.

Esta proposta é considerada um projeto de pesquisa aplicada e de inovação e pode ser trabalhada via subprojetos nos seguintes temas: Saúde; Energias Renováveis e Mudanças Climáticas; Óleo e Gás; Indústria 4.0; Educação; Cidades Inteligentes; entre outras que possam surgir dependendo dos bolsistas interessados na área de pesquisa.

Este projeto é a primeira versão na linha de Ciência de Dados (DS) ou Aprendizado de Máquina (ML) e Inteligência Artificial (IA) que apresento, o mesmo propõe pensar na ciência de dados não como um novo domínio de conhecimento a ser aprendido, mas como um novo conjunto de habilidades que pode se aplicar em qualquer área de especialização. Por termos a possibilidade de contar com alunos do curso de Bacharelado em Ciência da Computação, a proposta deste projeto se faz viável e trará benefícios para a comunidade universitária da UENF, no sentido de aprendizado teórico-prático em ciência de dados assim como na proposta de resultados baseada em modelos para decisões inteligentes.

### **Objetivo Geral**

- Estimular a realização de subprojetos em pesquisa aplicada e inovação que envolvem a Inteligência Artificial - IA, a Ciência de Dados (CD) e o Aprendizado de Máquinas (ML) no enfrentamento de desafios e demandas educacionais, sociais e até econômicas da comunidade fluminense, em especial na UENF, criando modelos e/ou metodologias para problemas específicos.

### **Objetivos Específicos**

Os objetivos específicos, para viabilizar a pesquisa, desenvolvimento e implementação de *pipelines* de dados, especialmente em subprojetos concretos que atendam demandas importantes; assim como o estímulo na construção de uma estrutura de aprendizado por projetos, para o desenvolvimento ou proposta de modelos que correspondam e ajudem na tomada de decisão inteligente, podem ser elencado na sequência:

- Estudar métodos analíticos e técnicas estatísticas, a fim de ajudar no processo de tomada de decisões.
- Pesquisar, projetar, implementar novas aplicações de análise de dados.
- Desenvolver experimentos, processos, ferramentas, sistemas e infraestrutura para suportar a manipulação de dados durante todo o seu ciclo de vida.
- Descobrir novas abordagens para gerar novos conhecimentos
- Definir e implementar estratégias de gerenciamento de dados para a coleta, armazenamento, preservação e disponibilização para futuro processamento;
- Estudar e manipular com destreza o editor de texto Latex, para digitação dos relatórios e trabalhos que surgiram a partir desta pesquisa inicial;

### **4. METODOLOGIA E ESTRATÉGIA DE AÇÃO**

A proposta de projeto “Ciência da Dados dirigido por Projetos” se fundamenta pela necessidade da formação de profissionais em Ciência da Computação com vasto conhecimento e prática em modelagem, análises e desenvolvimento de aplicações inteligentes com grande volume de dados. Este projeto de pesquisa, está sendo proposto para ser desenvolvido via subprojetos de interesse pessoal de cada bolsista, sempre que esteja dentro da linha de pesquisa. Na sequência descreve-se a metodologia e estratégia de ação que será utilizada para a realização dos subprojetos que assim sejam propostos nos diferentes planos de trabalho dos estudantes interessados na área. O foco principal deste projeto consiste em criar um aprendizado baseado em projetos usando, sempre que possível, os *datasets*

disponibilizados na plataforma Kaggle<sup>7</sup>. Lembrando que a escolha dos recursos e/ou projetos será realizado de acordo com a preferência e estudos realizados por cada um dos bolsistas vinculados ao projeto.

Para o desenvolvimento dos projetos sugeridos pelos estudantes interessados no tema “Ciência de Dados”, seguiremos o ciclo de uma aplicação de dados conforme os cinco passos definidos no livro “Principles of Data Science”(OZDEMIR, 2016) e elencado na sequência:

1. Definição de Escopo.
2. Coleta de dados.
3. Limpeza e análise exploratória dos dados coletados.
4. Modelagem dos dados.
5. Visualização de dados e apresentação de resultados para público não especializado.

A metodologia traçada poderá ser a seguinte:

- Estudo de conceitos, cenários e previsões para a Inteligência Artificial (IA), Aprendizado de Máquina (ML) e Ciência de Dados (CD);
- Estudo de conceitos básicos de Estatística, visto sua relação e importância refletida na Figura 2.
- Estudo e/ou treinamento de alguns métodos e ferramentas que nos auxiliam quando trabalhamos com Dados.
  - Ferramentas Básicas (Bibliotecas e SQL);
  - Data Wrangling/Descriptive Analysis;
  - Machine Learning;
  - Data Visualization;
  - Deployment
- Estudar sobre o conceito de Dados e Big Data, passando por Bancos de Dados não Relacionais (NOSQL).

A partir do conhecimento da área a ser analisada e do ferramental disponível, estrutura-se uma pergunta passível de resposta com dados; neste momento é preciso que os dados relevantes sejam coletados (o que pode ser por meio de Web Scraping<sup>8</sup>, requisição à APIs, ou organizado a partir de um banco de dados pré existente), a seguir são aplicados algoritmos de mineração de dados (*Data Mining*<sup>9</sup>) para compreender a natureza das variáveis e como elas se

---

<sup>7</sup> <https://www.kaggle.com/>

<sup>8</sup> [Data Science Skills: Web scraping using python | by Kerry Parker](#)

<sup>9</sup> [5 Data Mining Techniques Every Data Scientist Should Know | by Sara A. Metwalli](#)

relacionam, assim como técnicas de limpeza e padronização de dados (*Data Wrangling*<sup>10</sup>); uma vez organizado e mapeado o *dataset*, é possível estruturar um modelo que responda a pergunta inicial formulada, utilizando algoritmos de Aprendizado de Máquina e finalmente os resultados da análise são organizados para serem apresentados e visualizados.

Python poderia ser a linguagem selecionada para desenvolver esse projeto, por ser a principal linguagem utilizada na área, especialmente em função das suas bibliotecas: *Pandas*<sup>11</sup> para a manipulação de tabelas com dados heterogêneos na sua estrutura “*DataSets*”; *Numpy*<sup>12</sup>, que traz a velocidade da linguagem C e um vasto framework para manipulações numéricas, especialmente em álgebra linear; *SciKit-Learn*<sup>13</sup> que disponibiliza os principais algoritmos para aprendizado de máquina já implementados com vasto endosso referendado da comunidade científica, e por fim as bibliotecas de visualização de dados, plotagem de gráficos, *Seaborn*<sup>14</sup>, e *Matplotlib*<sup>15</sup>.

## 5. MATERIAIS E MÉTODOS

Para que as pesquisas sejam realizadas nos prazos estabelecidos e com um nível satisfatório, serão utilizados materiais e recursos disponíveis na instituição, oferecidos principalmente pelos laboratórios e bibliotecas. Em particular, serão utilizadas as instalações do Laboratório de Ciências Matemáticas (LCMAT) pertencente ao Centro de Ciências Tecnológicas (CCT) da Universidade Estadual do Norte Fluminense (LCMAT/CCT-UENF), que oferece PC e impressora interconectados em rede local que por sua vez possui acesso à rede da UENF e à Internet.

Estas máquinas estão equipadas com softwares para simulação, programação de alto desempenho e editoração de texto bastante difundidos na comunidade científica (por exemplo, MATLAB, Fortran, C/C++, Latex, Java, Linux, Python) e capacidade para instalação de novos softwares necessários à pesquisa. O laboratório também dispõe de espaço para reuniões e apresentação de seminários, equipado com retroprojetor ou datashow, que completa o suporte físico oferecido aos seus usuários.

Analisando e tomando como referência a pandemia do Covid19 que tivemos no ano de 2020, também estamos analisando a possibilidade de termos que usar a plataforma do Google Education (Classroom, MEET e o e-mail institucional) na qual a UENF participa com acesso a recursos ilimitados, para trabalhar de forma remota no projeto pretendido.

---

10

<https://www.elderresearch.com/blog/what-is-data-wrangling-and-why-does-it-take-so-long/#:~:text=Data%20wrangling%20is%20the%20process,20%25%20for%20exploration%20and%20modeling.>

11 <https://pandas.pydata.org/>

12 <https://numpy.org/>

13 <https://sklearn.org/>

14 <https://seaborn.pydata.org/>

15 <https://matplotlib.org/>

## **6. RESULTADOS ESPERADOS**

O resultado esperado deste projeto de pesquisa, através dos subprojetos desenvolvidos nesta linha de pesquisa, é de quatro tipos: i) desenvolvimento de produtos “Data-Driven”; ii) formação de recursos humanos; iii) divulgação através de participação em eventos; iv) consolidação de um grupo de pesquisa em computação na região Norte Fluminense. Melhor detalhamento de cada um desses tipos é dado a seguir:

- Os produtos “Data-Driven” resultantes deste projeto poderão ser:
  - APIs;
  - Modelos preditivos das mais diversas naturezas;
  - Sistemas de recomendação;
  - Aplicativos Inteligentes;
  - Relatórios analíticos assertivos;
- Formação recursos humanos:
  - Projetos de Iniciação Científica e Tecnológica em Ciência da Computação.
  - Ampliar o conhecimento dos bolsistas em relação à pipeline, e principalmente cultura de dados de modo a promover a literacidade científica moderna e o embasamento em evidência nas suas esferas de atuação;
  - Ensinar a escrever artigos/resumos/relatórios que sejam publicados nos canais correspondentes.
  - Formação de alunos no nível de graduação, para trabalhar em projetos envolvendo IA, agentes inteligentes, ciência de dados, mineração de dados e/ou textos e sua utilização no desenvolvimento de projetos reais.
  - Desenvolvendo a capacidade de pesquisa dos envolvidos;
- Consolidação de um grupo de pesquisa em Inteligência Computacional Aplicada numa região de baixo IDH, de forma a permitir aos mesmos interferir no sistema de inovação local.
  - Aprimorar a análise e capacidade crítica dos estudantes em relação a trabalhos científicos publicados;
  - Atrair o interesse dos gestores da instituição visando a utilização dos resultados obtidos na tomada de decisão.

## **7. PLANO DE TRABALHO**

Este projeto constitui a primeira versão na área da Ciência de Dados, Aprendizado de Máquinas e Inteligência Artificial, onde propomos o aprofundamento por parte do estudante e o desenvolvimento da área guiada por projetos que podem ajudar na tomada de decisão em problemas reais da nossa sociedade. Com isto quero dizer que o plano de trabalho para cada aluno interessado nesta linha de pesquisa, será ajustado e definido segundo interesse e motivação do próprio estudante.

## REFERÊNCIAS BIBLIOGRÁFICAS

- Blei, D. M. Smyth, P. (2017). *Science and data science*. PNAS, 114, 8689–8692. Disponível em [Science and data science](#). Acesso em 05/12/2020.
- Clavera , Walter V. (2019). *Introdução ao Aprendizado de Máquina (Machine Learning) para Iniciantes*. Repositório de Temas de Saúde Público. Disponível em [Aprendizado de Máquina \(Machine Learning\)](#). Acesso em 4/12/2020.
- Donoho, D. (2017). *50 years of Data Science*. Journal of Computational and Graphical Statistics, 26, 745–766.
- GRUS, Joel. (2015). *Data Science from Scratch*. [S. l.]: O'Reilly.
- Ivan de Souza Silva. (2004). *A importância da inteligência artificial e dos sistemas especialistas*. Congresso Brasileiro de Ensino de Engenharia - COBENGE. Disponível em [http://www.abenge.org.br/cobenge/arquivos/15/artigos/09\\_158.pdf](http://www.abenge.org.br/cobenge/arquivos/15/artigos/09_158.pdf). Acesso em 6/12/2020.
- OZDEMIR, Sinan. (2016). *Principles of Data Science*. [S. l.]: Packt.
- PRETI, Mariana. *AI: O que é Inteligência Artificial e porque os robôs têm-se tornado tão populares recentemente*. Disponível em <https://www.c2ti.com.br/blog/AI-O-que-e-Inteligencia-Artificial-e-porque-os-robos-tem-se-tornado-tao-populares-recentemente-Inovacao>. Acesso em 6/12/2020.
- Samuel, Arthur L. (1959). *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal of Research and Development. **44**: 206–226. CiteSeerX 10.1.1.368.2254. doi:10.1147/rd.441.0206.
- SCHUTT, Rachel; O'NEIL, Cathy. (2014). *Doing Data Science: Straight Talk From the Frontlines*. [S. l.]: O'Reilly.
- Schlemmer, E. ( 2001). *Projetos de Aprendizagem Baseados em Problemas: uma metodologia interacionista/construtivista para formação de comunidades em Ambientes Virtuais de Aprendizagem*. Revista Digital da CVA - Ricesu, ISSN 1519-8529. v.1, n.2.
- VANDERPLAS, Jake. (2017). *Python Data Science Handbook: Essential Tools for Working with Data*. [S. l.]: O'Reilly.