**APRENDIZADO DE MÁQUINA**

**Prof. Cristiano Carvalho**

**Projeto Final (60 pontos)**

A proposta deste trabalho é aplicar técnicas de aprendizado de máquina e extrair conhecimento de bases de dadospara identificar padrões que possam ser úteis para o apoio à tomada de decisão de algum problema social ou empresarial.

O processo de *Data Science* é iterativo, o que pressupõe o retorno a etapas anteriores do processo para ajustes caso necessário.

# ATIVIDADES

1. Cada grupo deverá ter até 5 alunos.
2. Cada grupo deverá escolher uma base de dados **(de livre escolha)**. Neste trabalho serão dadas algumas ideias de bases de dados disponíveis que podem ser utilizadas.
3. Cada grupo deverá, a partir da base de dados escolhida, aplicar o processo de *Data Science,* para extrair informação relevante sobre um problema diverso.
4. Poderá ser utilizada qualquer ferramenta de aprendizado de máquina***.***
5. **O grupo deverá:**
   1. Identificar e detalhar o problema que será abordado;
      1. Sugere-se que o problema possua escopo bem delimitado.
   2. Compreender os dados e como eles podem ser utilizados para resolver o problema;
   3. Realizar a seleção dos dados relevantes (registros e atributos);
   4. Enriquecer e melhorar os dados;
   5. Preparar os dados de acordo com os algoritmos;
   6. Aplicar algoritmos de *machine learning* de dados (pelo menos 2);
   7. Explicar o motivo do uso de cada técnica;
   8. Fazer análise dos resultados.
6. **Deve ser entregue:**
   1. Tabelas com os dados pré processados;
   2. Arquivos contendo os códigos utilizados;
   3. Documentação contendo:
      1. Análise dos dados;
      2. Relatórios dos algoritmos;
      3. Motivos do uso de cada técnica;
      4. Apresentação e análise dos resultados alcançados.variações
7. **Data de entrega:** até 15 dias após o término da disciplina.
8. De 6 a 12 páginas (se tiver problemas com essa delimitação basta sinalizar)

# SUGESTÕES DE BASES DE DADOS

1. O Portal da Transparência nos Recursos Públicos Federais possui uma seção de download de dados, em https://www.portaltransparencia.gov.br/download-de-dados, que tem por objetivo “facilitar a consulta das informações disponíveis no Portal da Transparência do Governo Federal e oferecer ao usuário uma forma rápida e prática de obter e armazenar os dados”. Os dados são separados por:
   1. Despesas – Gastos Diretos;
   2. Despesas – Transferências;
   3. Programas Sociais,
   4. Despesas – Por Função ou Subfunção;
   5. Receitas;
   6. Convênios;
   7. Servidores;
   8. Sanções;
   9. Beneficiados da Lei 10.559/02;
   10. Imóveis Funcionais.
2. O INEP – Instituto Nacional de Estudos e Pesquisas Anísio Teixeira disponibiliza diversos microdados já formatados sobre a situação da educação no Brasil. É uma base rica, volumosa, e bem segmentada. Os microdados estão disponíveis em: <http://portal.inep.gov.br/basica-levantamentos-acessar>. Os microdados acessíveis são:
   1. Enade (2004-2017);
   2. Censo Escolar (1995-2017);
   3. Censo da Educação Superior (1995-2017);
   4. Censo dos Profissionais do Magistério (2003);
   5. Saeb (Aneb/Prova Brasil) (1995, 1997, 1999, 2001, 2003, 2005, 2011, 2013);
   6. Enem (1998-2017);
   7. Provão (1997-2003);
   8. Pnera (2004);
   9. Prova Brasil (2007, 2009, 2011);
   10. Pesquisa de Ações Discriminatórias no Âmbito Escolar (2008).
3. O Tribunal Superior Eleitoral (TSE) disponibiliza as mais diversas informações desses candidatos. Técnicas de mineração de dados podem ser aplicadas para buscar identificar padrões nessa base, permitindo uma melhor compreensão da nossa democracia. As informações fornecidas, de forma livre e gratuita, pelo Tribunal Superior Eleitoral através do portal “Repositório de dados eleitorais” (<http://www.tse.jus.br/eleicoes/repositorio-de-dados-eleitorais>). Os dados disponíveis são:
   1. Eleitorado
   2. Candidatos
   3. Resultados
   4. Prestação de Contas
4. O Ipeadata (<http://www.ipeadata.gov.br/>) provê dados econômicos e financeiros do Brasil em séries anuais, mensais e diárias na mesma unidade monetária. A busca das séries estatísticas pode ser feita por "Palavras-chave" na sua descrição, "Temas" ou assuntos de interesse, "Fontes" de publicação ou pelo "Nível geográfico" para o qual é possível agregar os dados.
5. O Kaggle - Your Home for Data Science ([Kaggle](https://www.kaggle.com/)) é um sistema de desafios de *machine learning*, com diversas bases bem documentadas e problemas preestabelecidos para serem resolvidos. É uma ótima fonte de bases de dados e problemas de mineração que podem ser utilizados no trabalho.
6. O Data.World (<https://data.world/>) é uma Corporação de Interesse Público, que possui pelo menos 18 milhões de “*datasets”* disponíveis (alguns abertos e outros com acesso limitado). Os temas são separados por:
   1. [Health](https://data.world/datasets/health)
   2. [Education](https://data.world/datasets/education)
   3. [Finance](https://data.world/datasets/finance)
   4. [Government](https://data.world/datasets/government)
   5. [Public Safety](https://data.world/datasets/public-safety)
   6. [Agriculture](https://data.world/datasets/agriculture)
   7. [Weather](https://data.world/datasets/weather)
7. O grupo poderá trazer bases operacionais da sua empresa ou de qualquer lugar que julgar conveniente. Neste caso é desejável que a base possa ser disponibilizada para o professor, para permitir uma melhor avaliação do contexto de uso das técnicas de mineração. Se necessário os dados podem (e devem) ser anonimizados.