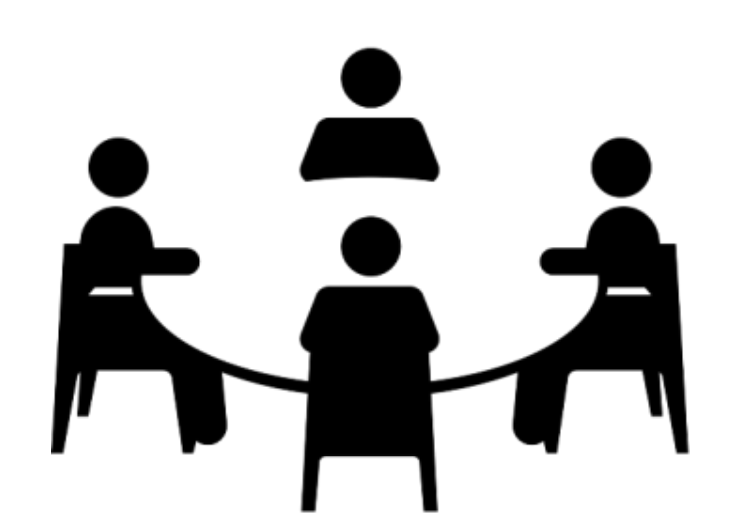
**Projeto A3 - UCs de Inteligência Artificial (IA) e Análise de Dados (AD).**

Este documento descreve o projeto da A3 para as UCs de Inteligência Artificial e Análise de Dados. Cada UC tem sua própria descrição de projeto e seu próprio conjunto de entregas. Para os alunos que cursam as duas UCs, algumas das entregas podem ser reaproveitadas, valendo para as duas UCs. A base de dados escolhida também pode ser utilizada nos conjuntos de entrega das duas UCs.

**Formação de Grupos:** Os grupos devem conter de **5 até 7** alunos.

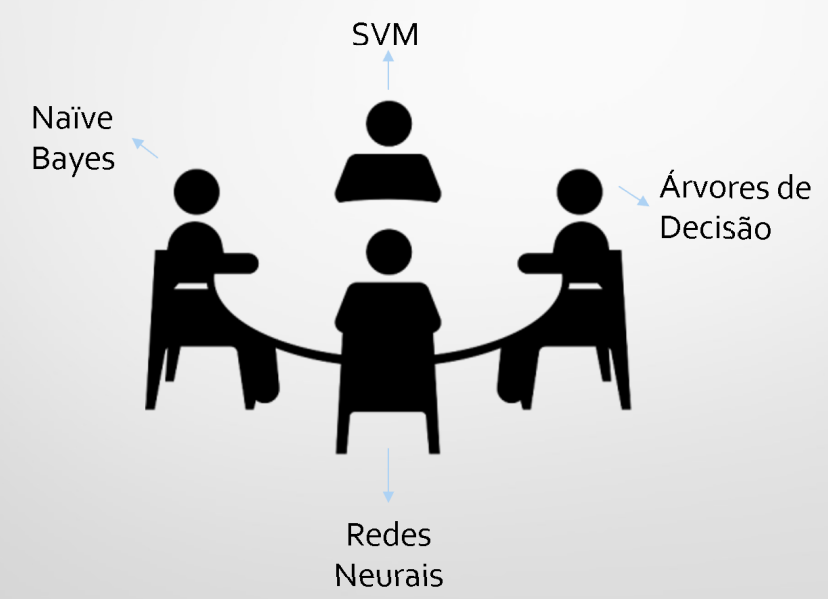
**O que será o projeto de IA?**

O projeto será fazer um Comitê de Classificadores, que consiste na comparação de diferentes técnicas de aprendizado de máquina para a resolução de um mesmo problema.



Cada grupo escolherá uma base de dados no **Kaggle** e aplicará métodos de IA para previsão de uma variável target. Depois de aplicar os métodos, deverá compará-los para chegar à conclusão de qual teve o melhor resultado, melhor custo benefício, etc.

**Kaggle:** https://www.kaggle.com/

****

**Quais métodos poderão ser aplicados?**

* Árvores de Decisão;
* Regressão Linear;
* K-Means;
* KNN – K-Nearest Neighbors;
* Naive Bayes;
* Redes Neurais;
* SVM – Suport Vector Machine;
* Qualquer outro método aprendido em aula.

Sendo que cada grupo deverá aplicar **no mínimo 3 métodos.**

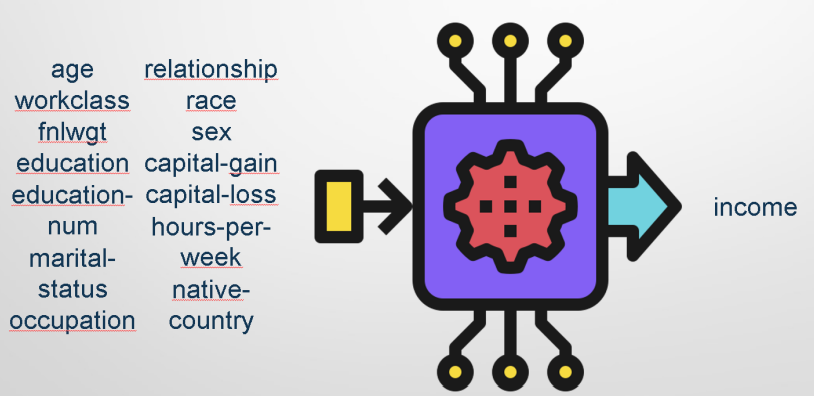
**Como chegar aos resultados?**

* Estudar a base de dados, entendendo quais colunas formam essa base e quais informações estão guardadas em cada coluna;
* O grupo deverá definir qual será a **variável target** que usará no projeto. A variável target será a variável de saída, a variável que se deseja “prever” algo, qual será a conclusão que se quer chegar;
* Aplicar os 3 métodos de IA escolhidos para previsão da variável target;
* Realizar a documentação, que deve incluir uma explicação da base de dados escolhida e sobre possíveis tratamentos de dados, além da comparação dos resultados obtidos pelos 3 métodos de IA utilizados.

Abaixo temos alguns exemplos de variável target para facilitar a escolha da base de dados:

**Exemplo 1:**

Podemos inferir (prever) a renda (income) com base nos demais atributos de uma base de dados, como mostrado na imagem:



**Exemplo 2:**

Podemos prever o preço de casas com base em suas características, como tamanho, número de quartos e idade. Considere uma tabela de dados formada pelas colunas abaixo:

* Coluna 1: Tamanho (metros quadrados);
* Coluna 2: Número de Quartos;
* Coluna 3: Idade da Casa (em anos);
* Coluna 4: Preço (em milhares de dólares).

Neste caso, a variável target é "Preço", que é uma variável contínua representando o preço de cada casa. O objetivo seria treinar um modelo de aprendizado de máquina que seja capaz de prever o preço de uma casa com base nas características fornecidas da mesma.

**Exemplo 3:**

Podemos determinar se um e-mail é spam ou não com base em palavras-chave no texto do e-mail. Considere uma tabela de dados formada pelas colunas abaixo:

* Coluna 1: Texto do Email;
* Coluna 2: Valor binário “É Spam?”.

Neste exemplo, a variável target é "É Spam?", que é uma variável binária. Ela assume o valor 1 se o e-mail for spam e 0 se não for spam. O modelo de aprendizado de máquina seria treinado para distinguir entre e-mails spam e não spam com base no texto do e-mail.

|  |  |
| --- | --- |
| **Texto do Email** | **É Spam?** |
| Oferta imperdível! Ganhe dinheiro rápido! | 1 |
| Reunião de equipe na Sexta-feira. | 0 |
| Ganhe um iPhone Grátis agora! | 1 |
| Confirme seu pedido de compra. | 0 |

**Entregas para a UC de Inteligência Artifical:**

As entregas deverão ser realizadas pelo Github até a data limite do cronograma.

|  |  |  |
| --- | --- | --- |
|  | Data | Descrição |
| Entrega 01 | 06/10 (Sexta) | Enviar um e-mail para o professor com as seguintes informações:   * Nome e RA de todos os integrantes do grupo; * Qual o campus em que estuda; * Para o projeto de qual UC o grupo foi formado; * Link do Github com o repositório criado pelo grupo para as futuras entregas.   Somente um aluno por grupo deve enviar o e-mail, representando todos os integrantes. Se o grupo será o mesmo para as duas UCs, somente um e-mail e um único repositório são suficientes. Se os grupos forem diferentes, dois repositórios e dois e-mails devem ser criados/enviados, um para cada UC. |
| Entrega 02 | 27/10 (Sexta) | Entregar a parte inicial da Documentação do projeto, incluindo: Explicação da Base de Dados: Os grupos devem apresentar a base de dados escolhida, descrevendo todas variáveis/colunas existentes na base, o que cada uma representa e quais os tipos de dados guardados em cada variável.  **Se o grupo estiver utilizando a mesma base de dados para as entregas das duas UCs (o que é recomendado), esta entrega é válida tanto para a Entrega 02 de IA quanto de AD.** Se bases de dados diferentes forem utilizadas, o grupo deve entregar uma documentação para cada uma das bases. |
| Entrega 03 | 03/11 (Sexta) | Definir a variável target, e atualizar a documentação incluindo:   * Texto explicando qual será a variável target (o que será previsto?) e justificando sua escolha. * Explicação sobre quais variáveis precisarão ser transformadas, caso existam (Análise Exploratória e Normalização) para aplicação das técnicas de IA. |
| Entrega 04 | 17/11 (Sexta) | * Entregar Código da Aplicação do Primeiro Método de IA; * Atualizar Documentação incluindo os resultados obtidos com o 1º método. |
| Entrega 05 | 24/11 (Sexta) | * Entregar Código da Aplicação do Segundo Método de IA; * Atualizar Documentação incluindo os resultados obtidos com o 2º método. |
| Entrega 06 | 28/11 (Terça) | * Entregar Código da Aplicação do Terceiro Método de IA; * Atualizar Documentação incluindo os resultados obtidos com o 3º método. * Atualizar Documentação contendo a comparação dos resultados obtidos por cada um dos métodos: Qual teve o melhor resultado e por que? |
| Pré-Banca | 28/11 (Terça) | Apresentação do projeto em aula. |

**O que será o projeto de Análise de Dados?**

Cada grupo escolherá uma base de dados no **Kaggle** e apresentará a **Estatística Descritiva** e **Inferencial** desta base. Os resultados devem ser apresentados por meio de **tabelas** e **gráficos**, utilizando a linguagem R, Python ou o software Power BI (Nas aulas de análise de dados veremos R e Power BI).

**Kaggle:** https://www.kaggle.com/

A Estatística Inferencial deve ser feita a partir do levantamento de uma hipótese, que será testada utilizando conceitos de **Teste de Hipótese** aprendidos em aula.

Os conceitos de Estatística Descritiva serão passados pelo professor Carmino, nas aulas online, enquanto os conceitos de Teste de Hipótese serão aprendidos na aula presencial.

Uma explicação e exemplos sobre Teste de Hipótese estão disponíveis no documento abaixo, porém este conceito será explicado com mais detalhes em aula em seu devido tempo.

[Guia - Inferência Estatística + Teste de Hipótese em R.docx](https://docs.google.com/document/d/1l-WXE-2tWXGtnCVK7TAWL10z7FpzhLFR/edit)

**Como chegar aos resultados?**

* Estudar a base de dados, entendendo quais colunas formam essa base e quais informações estão guardadas em cada coluna;
* O grupo deverá levantar uma hipótese, ou seja, devem definir uma afirmação que pode ou não ser verdadeira sobre a base de dados escolhida;
* Aplicar o Teste de Hipótese para verificar, matematicamente, se a hipótese levantada é verdadeira ou não para aquela base de dados;
* Realizar a documentação, que deve incluir: Uma explicação da base de dados escolhida e sobre possíveis tratamentos de dados, uma explicação sobre a Estatística Descritiva aplicada à base de dados e a apresentação dos resultados obtidos com o Teste de Hipótese.

**Entregas para a UC de Análise de Dados:**

As entregas deverão ser realizadas pelo Github até a data limite do cronograma.

|  |  |  |
| --- | --- | --- |
|  | Data | Descrição |
| Entrega 01 | 06/10 (Sexta) | Enviar um e-mail para o professor com as seguintes informações:   * Nome e RA de todos os integrantes do grupo; * Qual o campus em que estuda; * Para o projeto de qual UC o grupo foi formado; * Link do Github com o repositório criado pelo grupo para as futuras entregas.   Somente um aluno por grupo deve enviar o e-mail, representando todos os integrantes. Se o grupo será o mesmo para as UCs de IA e AD, somente um e-mail e um único repositório são suficientes. Se os grupos forem diferentes, dois repositórios e dois e-mails devem ser criados, um referente ao grupo de cada UC. |
| Entrega 02 | 27/10 (Sexta) | Entregar a parte inicial da Documentação do projeto, incluindo: Explicação da Base de Dados: Os grupos devem apresentar a base de dados escolhida, descrevendo todas variáveis/colunas existentes na base, o que cada uma representa e quais os tipos de dados guardados em cada variável.  **Se o grupo estiver utilizando a mesma base de dados para as entregas das duas UCs (o que é recomendado), esta entrega é válida tanto para a Entrega 02 de IA quanto de AD.** Se bases de dados diferentes forem utilizadas, o grupo deve entregar uma documentação para cada uma das bases. |
| Entrega 03 | 03/11 (Sexta) | Levantar a hipótese que será testada, e atualizar a documentação incluindo:   * Texto explicando qual hipótese será testada e justificando sua escolha. * Explicação sobre quais variáveis precisam ser transformadas, caso existam (Tratamento dos Dados) para aplicação dos cálculos de teste de hipótese. |
| Entrega 04 | 17/11 (Sexta) | Atualizar Documentação incluindo a Estatística Descritiva da base de dados escolhida. |
| Entrega 05 | 24/11 (Sexta) | Entregar Código da aplicação do Teste de Hipótese. |
| Entrega 06 | 01/12 (Sexta) | Entregar Documentação completa, incluindo apresentação dos resultados obtidos com o Teste de Hipótese em formato tabular e gráfico. |
| Pré-Banca | 01/12 (Sexta) | Apresentação do projeto em aula. |

**Instruções Gerais de Entrega, para as duas UCs.**

* A Entrega 01 deve ser realizada por e-mail: [evandro.ferraz@saojudas.br](mailto:evandro.ferraz@saojudas.br). Somente 1 aluno por grupo deve enviar o e-mail, representando todo seu grupo.
* As entregas subsequentes devem ser feitas atualizando o repositório do Github referente ao projeto.
* **A base de dados escolhida deve ser validada com o professor antes da Entrega 02 (27/10), em aula ou pelo email:** [**evandro.ferraz@saojudas.br**](mailto:evandro.ferraz@saojudas.br)**.**
* Para a UC de Inteligência Artificial: As entregas de Código devem ser feitas através de um arquivo em Python (.py ou .ipynb). As entregas da Documentação devem ser feitas a partir de um arquivo do Word (.docx) ou incluídas como bloco de texto em um arquivo notebook (.ipynb).
* Para a UC de Análise de Dados: A Estatística Descritiva deve ser entregue em formato de texto, a partir de um arquivo do Word (.docx) ou incluídas como bloco de texto em um arquivo notebook (.ipynb). O Teste de Hipótese da Estatística Inferencial deve ser programado utilizando R ou Python, e entregue a partir de um script em R ou um arquivo Python (.py ou .ipynb). Os resultados devem ser exibidos em formato tabular ou gráfico. Essa visualização dos dados pode ser gerada utilizando código (em Python ou R) ou com o software Power BI.

**Avaliação:** Para as duas UCs, a pontuação da A3 será considerada da seguinte forma:

* Entrega das etapas na data correta = Total de 10 pontos, sendo 2 pontos por cada entrega a partir da Entrega 02.
* Apresentação em aula (Pré-banca) = 10 pontos;
* Apresentação do projeto na Expo = 10 pontos;
* Total do projeto = 30 pontos.

Os outros 10 pontos da A3 (totalizando 40 pontos) de cada UC serão obtidos através da conclusão do Curso Dual respectivo. Informações sobre o Curso Dual estão disponíveis no Ulife, na aba “Material Complementar”, ou no Drive disponibilizado em aula, de cada uma das disciplinas. Qualquer dúvida podem perguntar em aula ou mandem no e-mail: [evandro.ferraz@saojudas.br](mailto:evandro.ferraz@saojudas.br) que eu respondo assim que puder.

**Exemplos de escolha de base de dados no Kaggle:**

Dica: Pesquisem “Machine Learning” na barra de pesquisa do Kaggle e escolham uma base de dados voltada para IA. Levantamento de Hipótese, que seria a entrega técnica de AD, é um conceito de aplicação mais simples.

**Exemplo 1:**

O conjunto de dados consiste em diversas variáveis preditoras médicas (variáveis independentes) e uma variável target (dependente): Outcome, que indica se um paciente possui diabetes ou não.

<https://www.kaggle.com/code/cdabakoglu/heart-disease-classifications-machine-learning>

**Variável Target:** Outcome: Essa variável indica se o paciente possui diabetes ou não, sendo 1 se o paciente possuir e 0 se não possuir.

**Possível Levantamento de Hipótese:** Pacientes com um número maior de IMC (Índice de Massa Corporal, representado pela coluna BMI) possuem uma chance mais alta de ter diabetes?

**Exemplo 2:**

O conjunto de dados consiste em diversas variáveis independentes, que representam informações contidas em exames de câncer de mama, e uma variável target (independente) que representa a categoria do tumor encontrado no exame. Nesta base de dados, considera-se apenas exames onde um tumor foi encontrado.

<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

**Variável Target:** Diagnóstico: A variável de previsão (target) é a variável que representa o resultado do diagnóstico do tumor no seio (coluna diagnosis), sendo M = maligno e B = benigno.

**Possível Levantamento de Hipótese:** Exames com números mais altos no desvio padrão de valores da escala de cinza (coluna texture\_mean) possuem uma tendência maior ao diagnóstico de tumores malignos?