


| | |
|---|--|
|  <small>Centro Universitário Farias Brito</small> | TRABALHO DA V2 |
| | CURSO: CIÊNCIA DA COMPUTAÇÃO DISCIPLINA: CIÊNCIA DE DADOS PERÍODO: 2024.2 – TURNO: NOITE PROFESSOR(A): CLEILTON LIMA ROCHA |
| | ALUNO(A): _____ |
| | |

Informações gerais:

- **Entregável:** Documento com explicações, Jupyter notebook e exportação do jupyter notebook em HTML.


Descrição

Sua empresa deseja contribuir com o bem-estar da sociedade e está engajada em desenvolver uma cultura de inovação aberta e inovação social, acerca disso ela deseja contribuir com ONGs protetoras de animais. Ela se prontificou a disponibilizar recursos para desenvolver uma solução que irá ajudar os biólogos **a classificar as espécies de pinguins e fornecer alguns insights para os biólogos** na identificação daqueles.

Dicionário de Dados

O conjunto de dados utilizados ainda é o mesmo dos pinguins, porém será considerado apenas as variáveis numéricas para construção dos modelos, logo elimine as variáveis que não são numéricas .

- **espécies (Variável alvo):** espécies de pinguins (Chinstrap, Adélie ou Gentoo)
- **culmen_length_mm:** comprimento do crista dorsal do bico das aves (mm)
- **culmen_depth_mm:** profundidade do crista dorsal do bico das aves (mm)
- **flipper_length_mm:** comprimento da nadadeira (mm)
- **body_mass_g:** massa corporal (g)
- ~~ilha: nome da ilha (Dream, Torgersen ou Biscoc) no Arquipélago Palmer (Antártica)~~
- ~~sexo: sexo de pinguim~~

| | |
|---|-----------------------------------|
|  <small>Centro Universitário Farias Brito</small> | TRABALHO DA V2 |
| | CURSO: CIÊNCIA DA COMPUTAÇÃO |
| | DISCIPLINA: CIÊNCIA DE DADOS |
| | PERÍODO: 2024.2 – TURNO: NOITE |
| | PROFESSOR(A): CLEILTON LIMA ROCHA |
| | ALUNO(A): _____ |

Atividades

1. Selecione duas soluções candidatas dentre ([KNN](#), [Árvore de Decisão](#) e [Floresta Aleatória](#)), justificando as suas escolhas.
2. Defina uma [métrica de classificação](#) (por exemplo, F1, Recall, Precisão, AUC ROC, ...) para analisar os resultados construídos através da [matriz de confusão](#), justificando sua escolha.
3. Crie um modelo de classificação para cada algoritmo selecionado e compare os resultados.
 - a. Durante o treinamento dos modelos aplique validação cruzada e grid search ou random searching.
 - b. **Se o melhor modelo** permitir você analisar a importância dos atributos, você listá-los (veja [feature importances](#)).
 - i. **Opcionalmente** retreine o modelo com as 2 ou 3 das melhores features e análise novamente os resultados
4. Verifique se o modelo está sofrendo com Underfitting ou Overfitting, justifique sua resposta.
5. Desenvolva uma análise de clusters (usando [K-means](#)), analisando o número de clusters e compare os resultados com os resultados encontrados em relação ao número de espécies existente (grupos e centróides), não esqueça de normalizar os valores.