**Atividade Prévia – Trabalho Prático**

|  |  |
| --- | --- |
| **Disciplina** | **SMA - Seleção de Modelos de Aprendizado de Máquina** |

**Objetivos**

Exercitar os seguintes conceitos vistos em aula:

* Exercitar conceitos sobre medidas de desempenho para regressão;
* Modelar um problema como uma tarefa de regressão;
* Avaliar um modelo de regressão;
* Exercitar conceitos sobre medidas de desempenho para classificação;
* Modelar um problema como uma tarefa de classificação;
* Avaliar um modelo de classificação;
* Exercitar conceitos sobre medidas de desempenho para clusterização;
* Modelar um problema como uma tarefa de clusterização;
* Avaliar um modelo de clustering.

**Enunciado**

Neste trabalho, vamos exercitar conceitos sobre medidas de desempenho vistas em sala de aula, a partir da modelagem de 3 problemas diferentes. Vamos modelar um problema de regressão, um de classificação e um de clusterização. Para o problema de regressão, o modelo de **regressão linear.** Para o de classificação, usaremos um **SVM.** Para o de clusterização vamos utilizar o algoritmo **kmeans.**

**Atividades**

Os alunos deverão desempenhar as seguintes atividades:

1. Selecionar TRÊS bases de dados diferentes para uma tarefa de regressão, uma tarefa de classificação e uma tarefa de clusterização. Atenção na seleção da base de dados para que o processamento de treinamento e teste de modelos seja viável. Possíveis fontes: UCI, Kaggle, OpenML.
2. Obter informações sobre números de features e instâncias dos datasets.
3. Identificar a existência de dados faltantes nos datasets.
4. Separar os conjuntos de treino e teste, usando a função **train\_test\_slipt**, com **test\_size = 0.35.**
5. Importar o sklearn para:
6. Aplicar à base **de regressão** o modelo de **regressão linear**.
7. Avaliar as métricas R2, MAE e MSE
8. Aplicar à base **classificação** o modelo **SVC, com kernel = rbf**.
9. Avaliar as métricas Acurácia, Precision, Recall, F1 e AUROC.
10. Aplicar à base **de clusterização** o modelo **kmeans**.
11. Identificar o número de clusters mais adequado de acordo com o dataset.
12. Avaliar as métricas Coeficiente de Silhueta, Davies-Bouldin Score e Mutual Information.

**Obs. Todas as sementes (random\_state) devem ser fixadas para avaliação da solução e das análises enviadas. Você deve enviar o notebook e o conjunto de dados selecionado.**