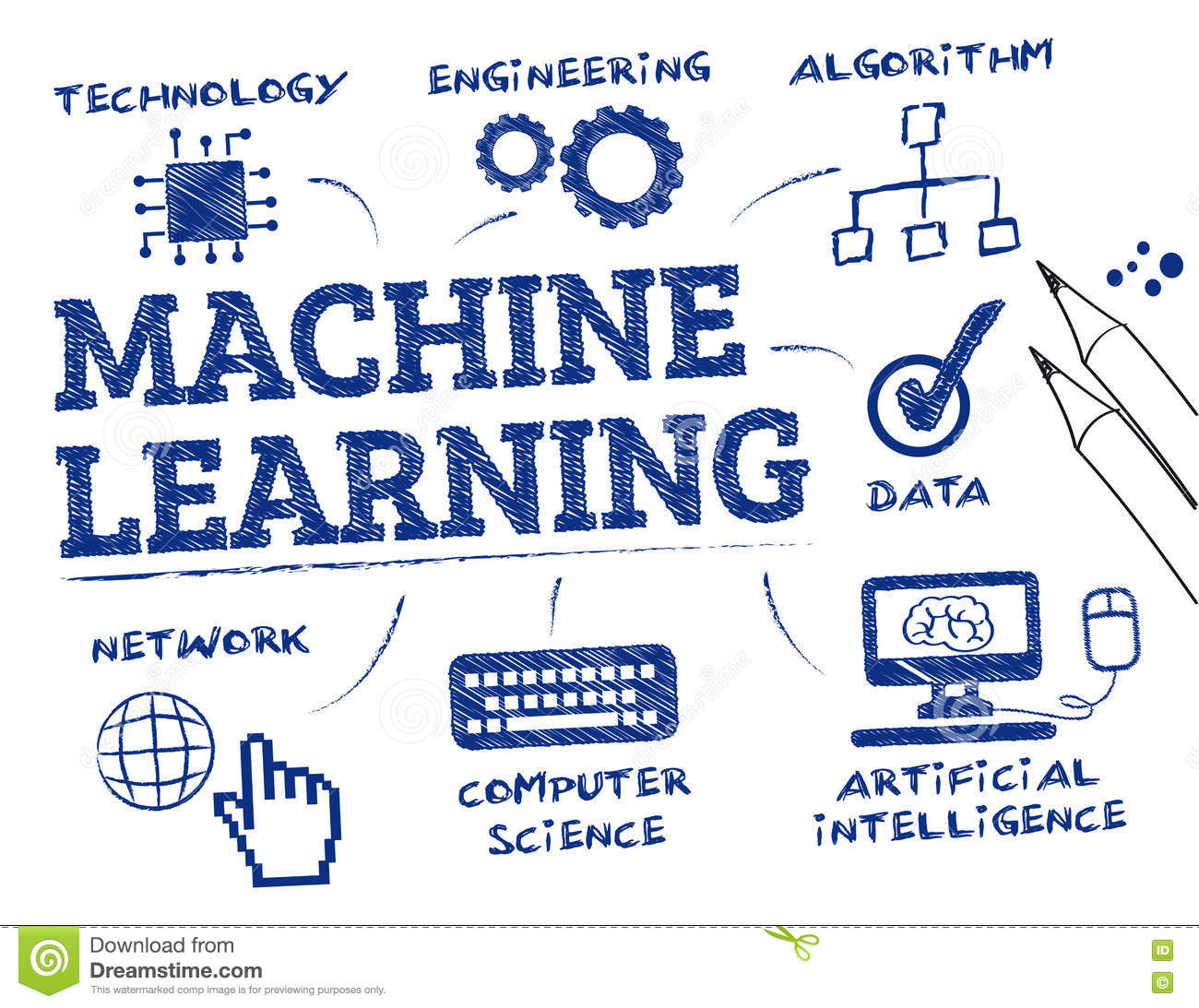
Engenharia do Conhecimento

**Projeto 2**



Grupo: 41

Francisco Henriques – 56348

Guilherme Marques – 55472

Miguel Seabra – 56344

Docentes: Sofia Teixeira, Cátia Pesquita

Data: 17/05/2023

* **Registo de horas de trabalho:**

|  |  |
| --- | --- |
| Aluno | Horas de Trabalho |
| Francisco Henriques  Guilherme Marques  Miguel Seabra | 22 Horas  0 Horas  0 Horas |

* **Introdução e Objetivos:**

O objetivo deste projeto baseia-se em desenvolver o melhor modelo de classificação possível para prever a variável *Biodegradable,* utilizando uma versão aumentada e editada do conjunto de dados: *QSAR biodegradation Data Set.* Disponibilizado pelos docentes da UC. A tarefa consistiu em 3 passos gerais como: Processar os dados disponibilizados; Examinar diferentes Hiper parâmetros para cada modelo; Selecionar as features mais importantes; Selecionar o melhor modelo;

Para atingir este objetivo, foi utilizado um JupyterNotebook de modo a ser mais fácil a execução do código necessário e da organização do output do mesmo. Esse ficheiro está em anexo junto deste relatório.

* **Processamento de dados:**

Tendo em conta que o conjunto de dados era relativamente grande (cerca de 42 colunas) e que não estava normalizado, existindo células em branco nos dados, começámos por carregar o csv disponibilizado para um panda, sendo depois utilizado o método KNNInputer para **preencher as células vazias**, este método tem em conta um K numero de vizinhos, em que neste caso foi escolhido K=5, de modo a não termos overfitting nem underfitting, com estes 5 vizinhos mais próximos o KNNInputer atribui um valor médio desses vizinhos à célula vazia.

Tendo o conjunto de dados completo e tendo em conta que a variável alvo (Biodegradable) era uma variável binária (RB ou NRB)