TASK 1 - PROBABILISTIC ARTIFICIAL INTELLIGENCE

The final solution of the task was implemented by computing the mean and the kernel of the posterior distribution from scratch, with the closed loop formula provided in the lecture. A 3D visualization of the dataset was very helpful to understand and handle the problem in a better way.

To cope with the computational issue of the problem, due to the very large dataset, inducing points method was implemented with a further Nyström approximation of the kernel function. The inducing points were selected by taking points equally spaced in the domain of the problem (from -1 to 1 for both features' dimensions).

For the model selection part, my final choice fell on the RBF kernel with a lengthscale of 0.05. The kernel type and the hyperparameters were tuned by evaluating the model with K-fold cross-validation (K=5) on the average score.

To handle the asymmetric cost, we set the prior mean to have a value between 0.5 and 1. In this way, when there is uncertainty about the prediction, we reduce the possibility of underestimating the pollution value. Optimal results were obtained by setting the prior mean to 0.8 (obtained via cross-validation).

A parallel solution was implemented with GPflow library, by means of the class SGPR, which performs Sparse GP Regression with inducing points method. Optimization of the hyperparameters over the marginal log likelihood was performed and similar results were obtained. However, I was not able to run the solution on Docker.