Crata AI - Data Challenge

**General Overview**

You have been hired by an investment firm to deliver an end-to-end AI solution that supports and facilitates their investment decision making process. In particular, the investment fund is specialised in investing in projects focused on sustainability and environmental impact.

In their selection process, the first thing the fund typically does is to analyse a company’s Linkedin profile (in particular their About and other textual components) in order to identify whether a given company is focused on sustainability goals. Up to now they have been manually going through and classifying linkedin pages as follows:  
  
**Label**

0 if not Sustainable

1 if Sustainable

Their manual labelling process has allowed them to create a dataset (attached) with linkedin profiles (and their textual data) and its classification towards whether the profiles adhere to sustainability or not. Now they have thought that they could actually automate this process by means of building a classification algorithm. As stated before, you have been hired to carry out this task.

The investment firm wants you to focus on three key aspects for the completion of your project:   
  
1. Initial Dataset Analysis and Insights

2. Training of a Classification Algorithm for identifying if a text is relating to sustainability

3. Model with API endpoint in production and simple example interface to use and showcase results

4. Guidelines and proposal for efficient data storage of both the provided dataset and any transformations/additions to the data carried out during the process of this task.

**Some Guidelines**

* You can provide the output in any format you like
* Ensure you select an appropriate classification algorithm and to justify why you selected it.
* We are looking both for code but also justification and reasoning of the steps you do, your analysis, the selection of the classification algorithm
* The API endpoint created should be executable in a local environment