Corpy&Co. Inc.		RECRUITMENT	Version	1.1
			Last modif.	2022-10-20
Branch	-	*ASSIGNMENT AI & ML ENGINEER*	Page	1 / 4
Path			Short ID	RCR001

Intended recipient	CANDIDATE
Sensitivity	**CONFIDENTIAL INTERNAL & CANDIDATE USE ONLY**

[INDEX]

Throughout this document, the date format is YYYY-MM-DD.

Intended recipient	CANDIDATE
Sensitivity	**CONFIDENTIAL INTERNAL & CANDIDATE USE ONLY**

INSTRUCTIONS

This assignment is divided into two sections. It aims at assessing the following competencies:

- data analysis (EDA),
- coding skills as AI engineer,
- ML model selection and training,
- results interpretation and presentation skills.

The table right below recapitulates the deliverables that need to be completed and sent back to the recruitment team before the deadline. In particular, the presentation that will be shown during the technical interview – if you pass the basics requirements – cannot be changed after the deadline.

DELIVERABLES	Archive .zip containing all the deliverables. Please name it `assignment-ai-[YOUR-NAME]-[YYYYMMDD].zip` with the date being the assignment deadline. The archive shall contain: 1. the present document as a reference, 2. a presentation (slides) in .pdf format, 3. the source code.
Note	The presentation should be written in English. You will split the presentation in two main sections, following the order of the questions. The presentation should ideally be between 20min and 30min long. The source code should not be more than 50MB. Please include a script that automatically downloads the required libraries, dependencies, the dataset and the checkpoints, or any other heavy files. Include a README and structure your project folder in a way that is easy to understand by the recruitment staff. Results should be reproductible.
QUESTION 1.	This is the first section of the presentation. Answer each question as one slide, you can add illustrations, histograms, and graphs along with text.
QUESTION 2.	This is the second section of the presentation. Describe the reasoning, the model description, and interpretation of the results. Please include graphs and visual analyses of the results. The code should be packaged as well. Please name the code folder `source-code-ai-[YOUR-NAME]-[YYYYMMDD]`

Send the assignment before the deadline that has been given to you.

Link to submission form	https://forms.gle/FGVpKozyqVs6msRa6

Archive name	`assignment-ai-[YOUR-NAME]-[YYYYMMDD].zip`
Archive content (deliverables)	**Name**
Assignment Sheet	`assignment-sheet-ai-[YOUR-NAME]-[YYYYMMDD].pdf`
Presentation	`presentation-ai-[YOUR-NAME]-[YYYYMMDD].pdf`
Source code	`source-code-ai-[YOUR-NAME]-[YYYYMMDD]`

QUESTION 1. DATA ANALYSIS

The purpose of this question is to evaluate the ability of the candidate to assess the properties and quality of a given dataset concerning a task at hand. Please answer each item below as one slide of the presentation that will be needed for the next interview.

The reviewer is expected to understand the following points throughout the content of the slides and your talk:

- What are your findings?
- Did you encounter any difficulties?
- How did you overcome those difficulties?
- What are the reasons behind the choices you made during the analysis ?

The dataset is MVTec Screws. This dataset is used for Anomaly Detection: some samples represent “good” items and other samples represent “anomalous” items. For your convenience, here is a link to the version that you are expected to use:
https://drive.google.com/file/d/11ozVs6zByFjs9viD3VIIP6qKFgjZwv9E/view?usp=sharing

The archive contains the following directories:

`train/good/`	training good samples
`train/not-good/`	training anomalous samples
`test/`	test images

Please notice that no categories are given for the test images. You are free to use them with manual qualitative inspection or to annotate them.

Provide an analysis on the qualitative and the quantitative aspects of the data:

a. How many samples are there?
b. Are the samples evenly distributed?
c. What kind of anomalies are there?
d. Is the data sufficient to train a model?
e. What kind of problems can we expect?
f. What can we do to improve the detection of anomalies in this data?
g. ***Optional:*** Anything else that you think is important (graphs, visualizations, measurements…)

QUESTION 2. CODING AND MODEL BUILDING

This question aims at evaluating the candidate's ability to build, train, and assess ML models. For this task, please use the data from QUESTION 1. What is expected is that the reviewer will understand the choice for the model you have made, the reason for the choices you have made, and the interpretation of the results. Please add all the answers as the second and last section of your presentation.

Use one ML model of your choice to build a binary classifier for the data provided in Assignment A. You may use Tensorflow, Pytorch, or anything else you desire. You may also use a pre-made model or a self-made model. Describe the model and explain the reasons for which you choose that model and framework.

Train your model on the provided data. Decide on how to split the data, the hyperparameters, the loss/evaluation functions, and describe why you have chosen these parameters.

Evaluate the performance of your model on the provided dataset. You are free to use the test data as an extra layer of evaluation and perform manual validation (qualitative and quantitative). Results should be reported using appropriate metrics. Please explain the metrics that you choose.

Independently of the quality of the performance of your model, please describe the reasons behind the results you obtain. Please reflect on any flaws in your method, or points that could have been done better.

We stress the importance of the reasoning and presentation / talking skills, despite the performance of your model being assessed as well.

Package your code inside the directory `source-code-ai-[YOUR-NAME]-[YYYYMMDD]`
A README should describe how to prepare the environment and run your code. Please do not include heavy files. Instead, use a script to automatically download them.

Optional: you can use `gdown` python library to automate the pulling of the dataset.