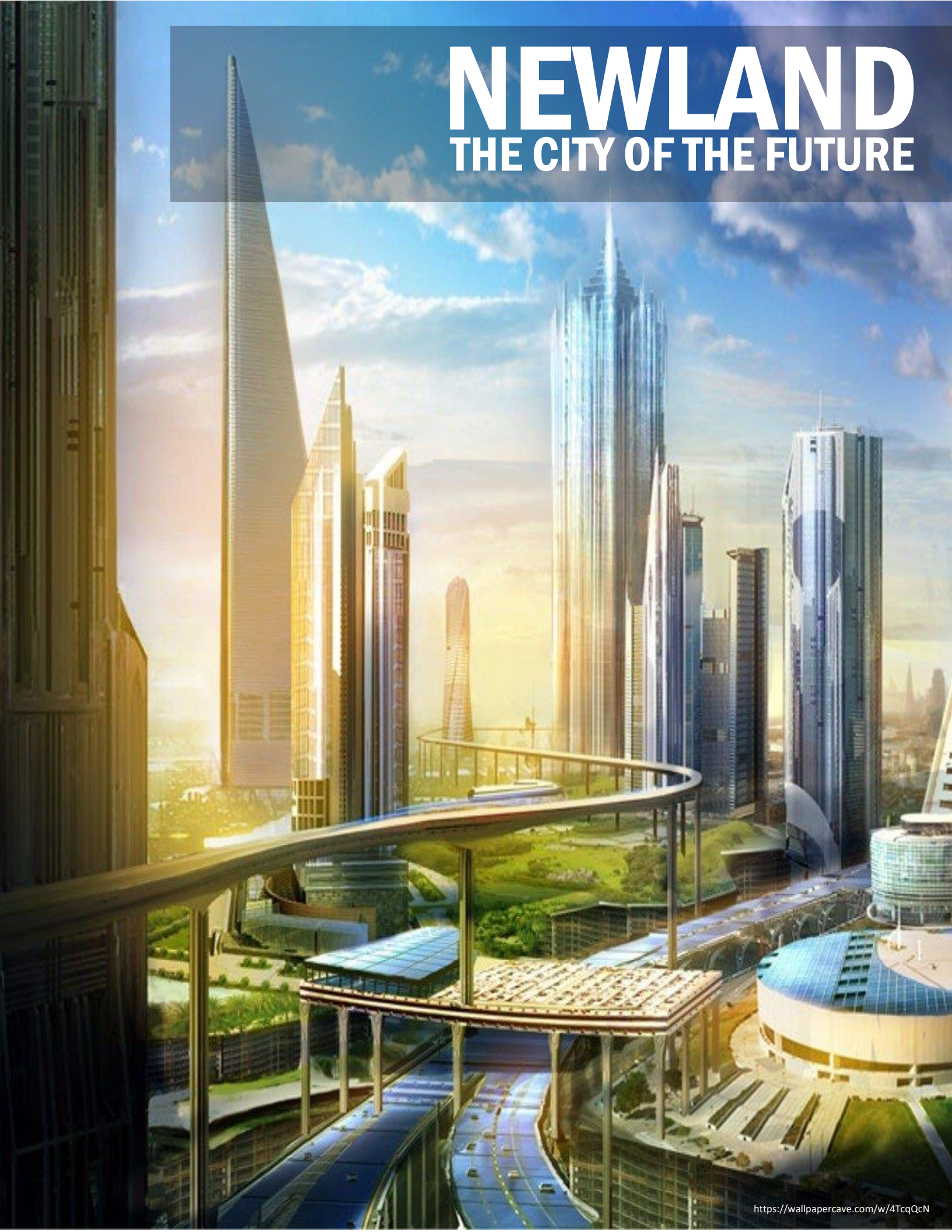
NEWLAND
THE CITY OF THE FUTURE
https://wallpapercave.com/w/4TcqQcN

## Objective of the document

The present document summarizes the guidelines for the Machine Learning project.

# Introduction

Welcome to Newland. In 2044, planet Earth was going through significant and accelerated climate changes that became unfeasible for all forms of life.

Five years earlier, a planet was discovered in our galaxy with conditions similar to those of planet Earth and a mission was developed to inhabit the new planet.

The mission, called "Newland", was launched in 2046, and sent 3 populated spaceships, each with a capacity of about 40,000 people, to the new planet. Unfortunately, in the first phase of the mission, 2 spaceships were reported missing.

The selection of the people to be sent was also controversial:

- The majority of the people (volunteers) were carefully chosen through an extensive selection process - Group A;
- More than 2500 people were payed to participate in the mission (those were people who have been considered by the state has essential people to have on an initial stage) – Group B;
- Around 1500 people payed to participate in the mission (those who were rejected in a selection process but entered by making a money offer) – Group C.

In 2048, and with another 100 new spaceships on the way (where the selection criteria remains controversial), the Newland government decided that residents should start paying taxes in order to make the new city more financially sustainable. They decided to apply a binary tax rate in which for people with an income below or equal to the average the rate would be 15% of income and for the remaining people 30% of their income. The first phase of this process is to identify the people who belong to each class. To this end, a first income analysis was made to 32500 people older than 17 years old, from which the government intends to create a predictive model to apply to people on their way to Newland.

Several groups of data scientists (including your group) were invited to create a predictive model based on a dataset of 22400 observations. This model will be applied to 10100 new observations (the test dataset) from which the government knows the value of income. The group of data scientists who obtain the best performing model will be invited to join the Newland government data science department.

# Evaluation Criteria

## Deliverables

- A jupyter notebook with all the needed code implemented to obtain the results presented in the report (***File naming format:*** GroupXX_ML202021_Predictive_Notebook.ipynb). This notebook should be the one who deliver the results defined as the selected solution in the Kaggle Platform.
- A report structured similarly to a scientific article (please check the document *report_structure.pdf* provided), that summarizes the analytical processes and the main conclusions obtained, with at most 6000 words (***File naming format:*** GroupXX_ML202021_Predictive_Report.pdf)
- An activity will be open on Kaggle where you should submit a csv file with the number of instances in the test set, containing the columns [Citizen_ID, Income], and only those columns. The Income column should contain the prediction (1 for "Higher than the Average", 0 for "Lower or equal to the average"). You should read the rules in this Kaggle activity page for further information.

## Guidelines for project evaluation

The project will be evaluated considering the following criteria:

- Model performance using F1 Score – Evaluation through Kaggle platform;
- The quality of the data exploration, pre-processing, modelling and assessment steps;
- A project that focus only on the techniques and methodologies approached during the practical classes will have at most 17 values. The remaining 3 values are possible to achieve if contributions based on self-study and creativity are applied, and clearly explained on the report.

### Note:

- All the topics mentioned will be evaluated based through the report - a well-structured and succinct report will have a big weight on the evaluation.
- The jupyter notebook will be analyzed if some doubt arises during the report evaluation or in the results in the Kaggle Platform.
- The report and the code will pass through a process of plagiarism checking.

## THE DATASET

| Variable | Description |
|---|---|
| Citizen_ID | Unique identifier of the citizen |
| Name | Name of the citizen (First name and surname) |
| Birthday | The date of Birth |
| Native Continent | The continent where the citizen belong in the planet Earth |
| Marital Status | The marital status of the citizen |
| Lives with | The household environment of the citizen |
| Base Area | The neighborhood of the citizen in Newland |
| Education Level | The education level of the citizen |
| Years of Education | The number of years of education of the citizen |
| Employment Sector | The employment sector of the citizen |
| Role | The job role of the citizen |
| Working Hours per week | The number of working hours per week of the citizen |
| Money Received | The money payed to the elements of Group B |
| Ticket Price | The money received by the elements of Group C |
| Income | The dependent variable (Where 1 is Income higher than the average and 0 Income Lower or equal to the average) |

The data has been split into two groups:

- Training set (22400 observations)

- Test set (10100 observations)

The training set should be used to build your machine learning models and assess the performance of it if needed. In this set, you also have the ground truth associated to each citizen, i.e., if the citizen has a higher income than the average or not.

The test set should be used to see how well your model performs on unseen data. In this set you don't have access to the ground truth, and the goal of your team is to predict that value (0 or 1) by using the model you created using the training set. The predicted values in the test set should be submitted on Kaggle.

The score of your predictions will be evaluated using F1 Score.