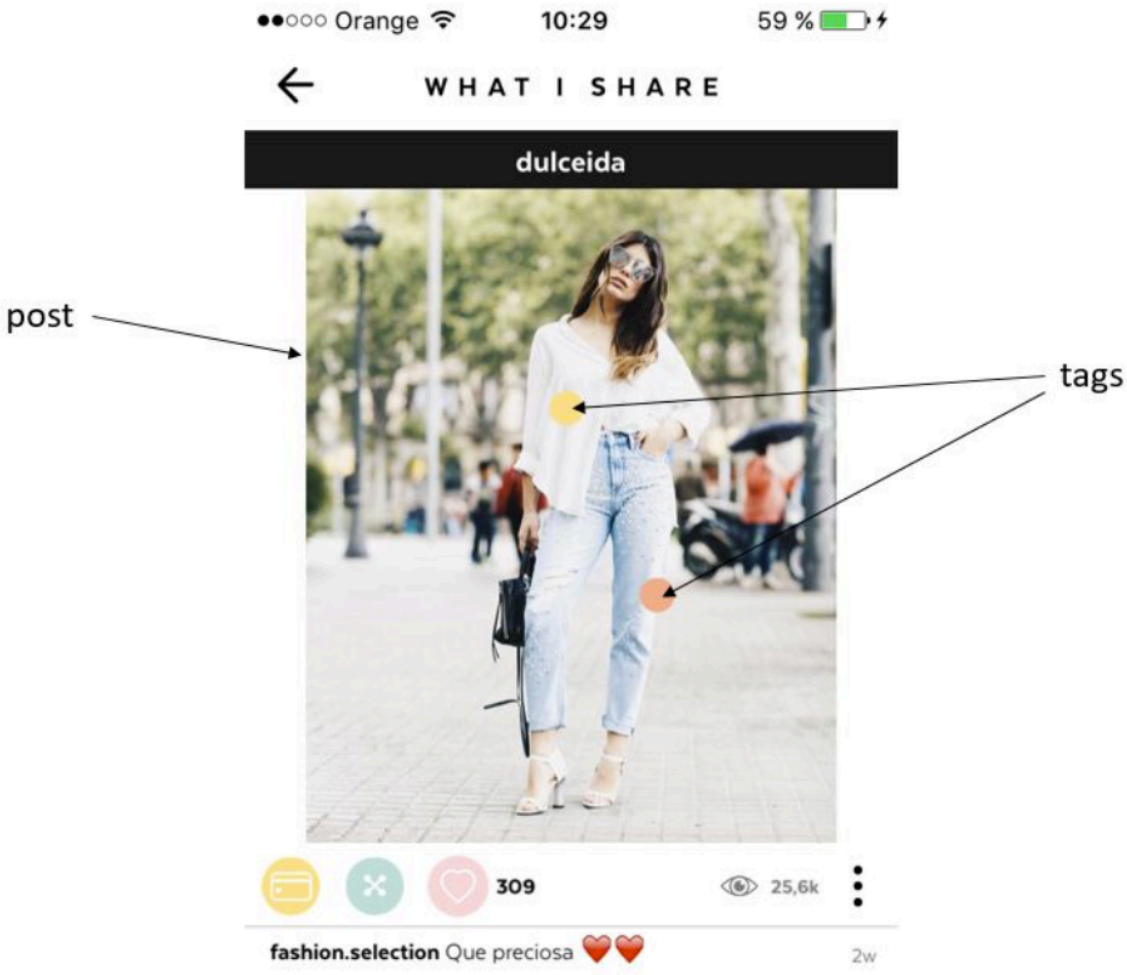
Machine Learning Assignment

This assignment is inspired by a social network where fashion lovers can upload and share their outfits with other users. Each item of clothing can be tagged specifying the brand, model, color, etc so you can keep track of what your friends or favourite influencers are wearing. Moreover, tags can be linked to the corresponding brand's e-commerce site. This way, if another user clicks on the tag and ends up buying the item, the original poster earns part of the purchase value.



The aim of the exercise is to predict the number of accumulated clicks a set of tags will have one month after their publication. Participants are given information about the owner of the post, the tags and the products linked to the tag.

Data model

There are three datasets to build the machine learning model: one for products (`products.csv`), one for users (`users.csv`) and the last one for tags (`tags_train.csv`).

products.csv

field	description
product_id	Product ID
product_info	Product information
description	Brief description of the product
brand_name	Product's brand name

users.csv

field	description
user_id	User ID
date_joined	Date of register of the user
country	The country of the user

tags_train.csv

field	description
tag_id	Tag ID
post_id	Post ID
product_id	Product ID
user_id	User ID
date_tag	Date of publication of the tag
color	Color of the tag
click_count	Accumulated click count of this tag one month after its publication. This is the target variable and obviously is only present in the train dataset.

Additionally, there is a `tags_test.csv` dataset that contains the tags for which the `click_count` you should predict.

Your answer

Your answer to the exercise should contain 3 parts:

- results csv file
- source code with model
- written description of the process, tools, challenges...

The results csv file should have this format:

```
tag_id,click_count
1,10
2,3
3,55
4,2503
```

The accuracy of prediction will be measured with the root mean squared error (RMSE) (square root of the mean squared difference between the predicted count of clicks and the actual count of clicks for each tag in the `tags_test.csv`)

You should attach source code you used to solve the assignment, and references to external dependencies like versions of ML libraries, tools, etc. Python scripts, Jupyter notebooks, R files or whatever you used is fine.

Also, we expect a written description of the steps you took in solving this exercise: what tools did you use? what challenges did you find? what other approaches would have been reasonable? why did you pick this one?