**UNIVERSITAT POLITÈCNICA DE CATALUNYA**

FACULTAT D'INFORMÀTICA DE BARCELONA

# Understanding salary using Machine Learning

Machine Learning

Spring 2022

Authors:
**Lorencio Abril, Jose Antonio**, *email*: jose.antonio.lorencio@estudiantat.upc.edu
**Mayorga Llano, Mariana**, *email*: mariana.mayorga@estudiantat.upc.edu

Professor: **Coma Puig, Bernat**

# Proposal

## Dataset

The original dataset used for this project is the Adult dataset [KB], which was extracted from the 1994 Census database by Ronny Kohavi and Barry Becker. It contains **14 features** describing demographic information about individuals, as well as a binary class based on income levels, using a threshold of $50,000 per year. The dataset includes **32561 instances**[1], each with 14 features, including information about age, workclass, education level, marital status, occupation, relationship, race, sex, native country, and more.

It is important to mention that the records of this set have already been filtered using the following conditions[2]: ((AAGE>16) and (AGI>100) and (AFNLWGT>1) and (HRSWK>0)). However, further pre-processing and cleaning will be done for this project.

### Attributes

From the attributes included in this data-set, 6 of them are represented as numerical attributes and the remaining 8 fields belong to a categorical feature. The class is a binary property describing whether an individual's income level is above or below $50,000 per year. The attributes are shown in the table below.

| Attribute | Type |
|---|---|
| age | continuous |
| workclass | discrete (8 categories) |
| fnlwgt | continuous |
| education | discrete (16 categories) |
| education-num | continuous |
| marital-status | discrete (7 categories) |
| occupation | discrete (14 categories) |

| Attribute | Type |
|---|---|
| relationship | discrete (6 categories) |
| race | discrete (5 categories) |
| sex | discrete (2 categories) |
| capital-gain | continuous |
| capital-loss | continuous |
| hours-per-week | continuous |
| native-Country | discrete (41 categories) |

| Class | Type |
|---|---|
| Income | discrete (binary) |

## Project Objective

1. The first and most important objective of the project is to develop models that, given demographic characteristics, can predict if a subject earns more or less than $50000 per

[1] In the site it says it has more than 40000, but we have checked it and the correct value is 32561.

[2] Basically, they extracted relevant records from the census. For instance, people under 16 are not earning any salary.

year.

2. Apart from this, given the nature of the dataset (e.g. there is unbalanced in the races representation or double as men as women), we plan to tackle classification fairness and bias, applying different methodologies that exist to deal with this.

   For this, we want to apply a similar approach to that in [HPS16], where the authors propose different methods to reduce the prediction biases of our models. This could be useful, for example, if our model was part of a bigger system, in which it is important for us not to be biased against a particular group.

3. Finally, we are interested in the approach to reduce bias towards Learned Latent Structure, as presinproceedings in [Ami+19]. We are not completely sure if this approach will be applicable in our project, because the dataset is quite modest, so we assess this as a possibility, rather than an objective.

# Bibliography


[Ami+19] Alexander Amini et al. "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure". In: Jan. 2019, pp. 289–295. DOI: 10.1145/3306618.3314243.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. *Equality of Opportunity in Supervised Learning*. 2016. arXiv: 1610.02413 [cs.LG].

[KB] Ronny Kohavi and Barry Becker. *Adult Data Set*. url: https://archive.ics.uci.edu/ml/datasets/Adult