Write the title of your report here

John Smith[1]*, Jennie Smith[1]

Abstract

Keywords
Optics — Interference — Diffraction

[1] *Department of Physics, Umeå University, Umeå, Sweden*
***Corresponding author**: john@smith.com
***Supervisor**: joe@doe.com

Contents

1. Introduction

2. Data analysis

2.1 Dataset

The dataset we decided to study is a labeled income prediction dataset. This dataset includes 14 features with information about the people in the srudy and a label with the income as either more than 50 000$ per year or less than or equal to 50 000 $ per year. This means that we are looking at a binary classification problem. A lot of the features are discrete where only a set number of options available. This includes features such as marital status, education and working class. The dataset features around 32500 data points.

2.2 Data cleaning and feature engineering

There were a couple of things with our dataset that had to be modified in order for it to be usable in our ML application. We find that some of the features are redundant or not interesting in our project. We romove the redundant feature education since there is another already numerically encoded feature containing the same data. We also chose to remove the feature 'fnlwgt' since it is a already calculated number that is used by the Census Bureau to estimate population statistics. Since we want to estimate the population statistics based on the other features and not the already calculated weight we remove this feature. We have a mix of numerical and non-numerical features in our dataset. Since the machine learning models cannot use non-numerical data we have to encode the non-numercial data into corresponding numbers. This is with the label encoder built into sci-kit learn and used on all non-numerical data.

2.3 Handling missing values

With our numerical version of the dataset we found with the info function in pandas that around 2500 values were NaN values. We reasoned that filling these values with something as the mean of the category does not make very much sense for our application. Since there are many discrete categories a mean value means nothing. Especially since we gave many categories arbitrary numbers

3. Model selection

4. Model Training and Hyperparameter Tuning

5. Model Evaluations

6.

References

[1] Steinhaus, H., Mathematical Snapshots, 3rd Edition. New York: Dover, pp. 93-94, (1999)

[2] Greivenkamp, J. E., Field Guide to Geometrical Optics, SPIE Press, Bellingham, WA, (2004)

[3] Pedrotti, F.L. and Pedrotti, L.S., Introduction to Optics, 3rd Edition, Addison-Wesley, (2006)

[4] UC Davis ChemWiki, Propagation of Error, Available at: https://chem.libretexts.org/Textbook_Maps/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Quantifying_Nature/

Significant_Digits/Propagation_of_Error, (Accessed: 10th March 2016).