

# Assignment 2

Start Assignment

- Due 18 Nov by 8:00
- Points 0
- Submitting a file upload
- Available after 6 Nov at 17:00

## Purpose

This assignment should be solved in groups of 3-4 students.


The overall purpose is to get familiar with standard machine learning tasks, specifically classification and regression.

The important parts of this assignment are reasoning and understanding the concepts. That means that **it does not matter if you get bad results** from the models **as long as you can understand and reason why it happens**. Spend time on understanding the data and process, and writing good reasoning, instead of trying to blindly optimize the models for the best score.

---

## Description

### Classification


The dataset titled [adult.csv \(https://ju.instructure.com/courses/12258/files/1850040?wrap=1\)](https://ju.instructure.com/courses/12258/files/1850040?wrap=1)  ([https://ju.instructure.com/courses/12258/files/1850040/download?download\\_frd=1](https://ju.instructure.com/courses/12258/files/1850040/download?download_frd=1)) gives information on the attributes of a group of people. Your task is to train a classification model to predict whether their individual income is  $\leq 50k$  or  $> 50k$ . Please perform the following tasks:

1. Conduct data pre-processing and feature selection.
2. Try at least three different machine learning classifiers.
3. Compare the classifiers by using a 10-fold cross-validation based on two different criteria (e.g. accuracy and AUC for the classification task).

4. Choose one of the machine learning classifiers from sub-task 2 that you think is best. Optimize one of the parameters of the machine learning classifier based on the criteria in sub-task 3.
5. With the selected machine learning classifier and optimized parameters, train the classifier with the chosen parameters on the entire dataset and save the predictor model.

The dataset is available here: [adult.csv \(https://ju.instructure.com/courses/12258/files/1850040?wrap=1\)](https://ju.instructure.com/courses/12258/files/1850040?wrap=1)  [\(https://ju.instructure.com/courses/12258/files/1850040/download?download\\_frd=1\)](https://ju.instructure.com/courses/12258/files/1850040/download?download_frd=1)

## Regression

The dataset titled [Housing.csv](https://www.kaggle.com/datasets/camnugent/california-housing-prices)  [\(https://www.kaggle.com/datasets/camnugent/california-housing-prices\)](https://www.kaggle.com/datasets/camnugent/california-housing-prices) gives attributes about houses in California. Your task is to train a regression model to predict the **medianHouseValue** (Median house value for households within a block (measured in US Dollars) for each house. Please, perform the following tasks:

1. Conduct data pre-processing and feature selection.
2. Try at least three different machine-learning regressors.
3. Compare the regressors by using a 10-fold cross-validation based on two different criteria (e.g. MSE and  $R^2$  value).
4. Choose one of the machine learning regressors from sub-task 2 that you think is best. Optimize one of the parameters of the machine learning regressor based on the criteria in sub-task 3.
5. With the selected machine learning regressor and optimized parameters, train the regressor with the chosen parameters on the entire dataset and save the predictor model.

The dataset is available here: [California housing](https://www.kaggle.com/datasets/camnugent/california-housing-prices)  [\(https://www.kaggle.com/datasets/camnugent/california-housing-prices\)](https://www.kaggle.com/datasets/camnugent/california-housing-prices)

---

## Submission Details

When submitting your solution, please use the Assignment group you belong to. To pass this assignment you should write a short text about your chosen approach. The exact format of your text is free, you do not need to mimic a scientific paper.

## The Report Structure

Your text must include at least these parts:

- A part where you reason about your pre-processing and feature selection.
  - Illustrations and visualizations of the data should be included.
- A part where you reason about your choice of machine learning model in classification and regression, and how it compares to the other models for the problem.
  - A short description of the parameter optimization of the chosen model should be included.
- A part about evaluation results from the classification and regression (i.e. charts, tables) that support your choice and reasoning for which model is chosen, along with a description of your evaluation.
  - Appropriate tables and graphs of the results should be included

## The Report Text

Your text **must** follow these guidelines:

- The text must be in English.
- The report cannot be more than 2000 words.
- The report must contain at least 4 figures or charts.
  - 2 figures or charts should depict the comparison between the three different classifiers and regressors.
  - The other figures and/or charts should depict the data, statistics or other visualizations.

## Passing

This assignment is graded using only Pass or Fail. To pass, your text must be well-structured, well-written and it should of course treat the chosen subject satisfactorily. To pass the assignments in the course (to get the points in Ladok), you must pass this assignment, along with the other remaining assignments in the course. See the course PM for further details on re-examination.

Submit your text in Canvas before the deadline as communicated through Canvas.

Questions about this assignment should be sent to: [Helena.lofstrom@ju.se](mailto:Helena.lofstrom@ju.se) (<mailto:helena.lofstrom@ju.se>) or [Cecilia.sonstrod@ju.se](mailto:Cecilia.sonstrod@ju.se) (<mailto:cecilia.sonstrod@ju.se>)

