

CSCI 335 Sections 2 and 3

Homework 2

Programming Assignment (16 points)

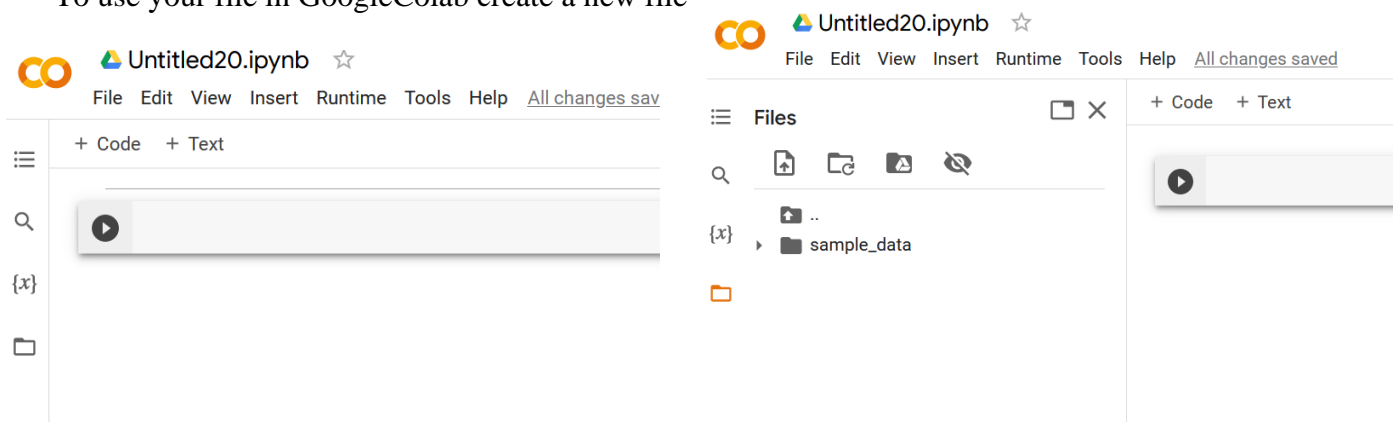
Predict the `median_house_value`. The file `houses.csv` contains Information about houses in Californian districts. Write a Python program (in jupyter notebook) that does the following:

- (1 point) Load the dataset.
- (2 point) Asses the data and clean them.
- (1 point) Create features X and targets y .
- (1 point) Divide the dataset into training, validation, and testing sets: (X_{train} , y_{train}), (X_{val} , y_{val}), and (X_{test} , y_{test})
- (1 point) Create a transformation (`ColumnTransformer`) that applies `OneHotEncoder` to the nominal column only (`remainder='passthrough'`). Fit it with X_{train} and print the first five lines of the transformed X_{train} .
- (3 points) Train `KNeighborsRegressor` on the transformed X_{train} and select the value of k on the transformed X_{val} . Choose a metric that you think evaluates this problem the best. Please, state why you chose it.
- (1 point) Report the results using X_{test} and an appropriate metric.
- (1 point) Modify the transformation applying a scaler to all numerical columns
- (3 point) Repeat the training with the scaled data.
- (2 point) Discuss the results. I would suggest plotting histograms of the errors for both cases. If the results are different, can you think of the reason?

Writing Assignment (4 points) Use Markdown/Text cell in your jupyter notebook

- (2 point) What is the difference between supervised and unsupervised problems?
- (2 point) What is the purpose of splitting data into training, validation, and testing sets? What is a typical split-up?

To use your file in GoogleColab create a new file



Click on the folder sign
You can also mount your Google Drive in Colab.

Drag your file in the area under the folders or click Upload