Objective :

Apply a machine learning algorithm and a neural network algorithm on a real life dataset.

Data :

You have 3 data sets at your disposition. You have to choose one dataset out of the 3 or you can work with your own data set if you want to work on a particular subjects :

diabetes.csv :

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The dataset consists of several medical predictor variables and one target variable, `Outcome`. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

Output variable :

Outcome (0 or 1)

house.csv :

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. It's a great dataset for evaluating simple regression models for predicting the price of a house depending on its characteristics.

Output variable :

price (continuous)

wine.csv :

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

Use machine learning to determine which physiochemical properties make a wine 'good'!

Output variable :

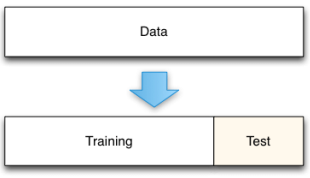
quality (score between 0 and 10)

Methodology :

Preprocess and load data :

Load the dataset that you have chosen to work with.
One of the first steps of the exploratory data process when the ultimate purpose is to predict the output, is to create visualizations that help get knowledge of the output and then to uncover relationships between the attributes and the output.
You can use multiple data visualization tools like barplots, histograms, boxplot, correlation matrix, pairs plot, PCA analysis.
Choose the variables you think will be useful for your prediction.
Take care of eliminating correlated features.
Do not hesitate to create new features as a combination of other features.
If you think they have a negative impact on your predictions, remove some outliers.

Engineering dataset :



Split your dataset into 2 parts :

- training set
- test set

If you think it is useful, apply standardization on your dataset. Be careful to apply the reverse transformation on your predictions at the end to compare your result with the true output.

Define model :

Choose a machine learning algorithm and fit it to the data.
Fit a neural network to the data.

For both models, try to use grid search to optimize the hyperparameters.

Evaluate models :

Evaluate each model on the test set and keep the best model.
Do a visualization of some predictions and compare it to the true predictions with the best model.

How to get a good grade ? 

- Comment every block of code and justify every choice. A visualization that is not commented is considered as not done.
  ex : The box plot shows 3 outliers, we remove it from our dataset for the rest of the study.
- It is more important to get bad predictions and to get a good methodology and good justification than the opposite.

Tips : 

- Use the python library sklearn for machine learning : https://scikit-learn.org/stable/
- Use the last notebook “Correction Neural networks practical lab.ipynb” to help yourself. The methodology that you must follow is very similar to this notebook.

Working environment :

You will work with google colab notebooks. Once you have done the project, download the notebook in .PDF format and you can share it with me at my mail address : kattin.dassance@hupi.fr Take care of running each cell before you share it with me so I don’t have to run your code again. So I want a PDF format file please with each line code running before in order to see all logs and prints.

The deadline is 2023/03/20

If you have any questions do not hesitate to contact me at : kattin.dassance@hupi.fr

Good luck ! :)