=> Machine Learning Assignment 2 <=
50 MARKS

Deadline

Submit all tasks on RuConnected by the **deadline** else default penalty per hour**.**

Task 0: Upvote the kaggle notebook [10 Marks]

Vote the kaggle notebook called **Lecture___iris_DL_end-to-end** and I will check it against your username/email address on Kaggle.

Task 1: Cleaning your data [5 + 5 = 10 Marks]

Pandas is a great tool for loading data from CSV files, containing dataframes(s) of pre-extracted features. However, some of these datasets include poorly formatted or even missing values.

1) Use what you have learnt during the Data Acquisition Cleanup step to handle missing data using the **median** strategy.
2) Print the number of **NaN** values **per feature.**
3) Moreover, **fix the order of one** of the steps in the program such that it resembles Best Practices.

The dataset is called *winequality-red.csv*, which is called in the *cleaned.py* Python script.

Deliverables:

- Modified Python script

Task 2: Scaling your data [2 + 3 + 5 + 5 + 5 = 20 Marks]

Not all classifiers benefit from data scaling, but a distance-based classifier like k-NN almost requires scaling by default.

1) Use the k-NN classifier on *winequality-red.csv* with **one** neighbour.
2) Scale the data,
3) Perform 5-fold cross validation (CV) on **70%** of the data and comment which scaler achieves the best **accuracy** and **f1-score (weighted)** on the CV data. Use what you have learnt to do so.
4) Include a **classification report** on the remaining **30% test data** using the same methods you used on the 5-fold CV.
5) Assign the classes to three bins to create a new dataframe. Now retrain and test as above to produce a new **classification report**.

Recommended to modify Task 1, and name the new script *cleaned_and_scaled.py*.

Deliverables:

- Modified Python script with comments to explain what/why you think works best.

Task 3: Making a medical diagnosis [25 marks]

In Task 5 of Assignment 1, you used the Breast Cancer dataset. Again make use of the built-in Scikit Breast Cancer dataset and convert it to a Pandas dataframe. Use all that you have learnt (and more if you want) to write a python script that aims to achieve the top f1-score on both k-NN and decision trees classifier.

You must follow Best Practices. Use comments to explain your reasoning for using particular EDA, feature processing etc.; the concepts learnt in **iris_DL_end-to-end** notebook will be helpful to conduct this task. **Hint:** Do not overkill with unnecessary EDA that does not allow for meaningful analyses.

Finally, note that **iris_DL_end-to-end** does not follow *Best Practices* perfectly; fix that.

Deliverables:

- Python script with comments to explain your 'decisions' based on EDA and other methods.