Compute Science Department
CS672 – Introduction to Deep Learning (CRN# 74071)
Fall 2023

Project #1 / Due 23-Oct-2023
This exercise will guide you through two important tasks in any Machine Learning / Deep Learning engagement:

- ➢ **Perform Exploratory Data Analysis**
- ➢ **Device a ML/DL Model (Classification Analysis)**

Performing Exploratory Data Analysis (EDA) on data is of paramount important for every Data Scientist / Data Analyst. Exploratory Data Analysis is often used to uncover various patterns present in the data and to draw conclusions from it. EDA is the core part when it comes to developing a Machine Learning model. This takes place through analysis and visualization of the data which will be fed to the Machine Learning Model. A Machine Learning Model is as good as the training data - you must understand it if you want to understand your model.

Prior commencing your efforts on coding, you must install the following libraries:

- pip install -q tensorflow_data_validation [visualization] (**)
  - https://pypi.org/project/tensorflow-data-validation/
- pip install apache-beam [interactive]
  - https://beam.apache.org/get-started/quickstart-py/
  - https://pypi.org/project/apache-beam/
- Install the GraphViz library
  - https://www.graphviz.org/download/

(A)
Perform an **explanatory data analysis** (**EDA**) on **Wine Class Classification** dataset donated to the Data Science community on June of 1991! Besides the need to build a model based on the data provided, you are asked to look for issues in the data and find correlation among the various variables in order to improve wine class classifications.

Write **Python** (or on any language contained within a notebook) scripts in order to complete the following tasks along with their output. All work should be done and submitted in a single **Notebook (Jupyter or Colab).**
1) Prep the data in order to be ready to be fed to a model.
Look for missing, null, NaN records.
Find outliers.
Transform data – all entries should be numeric.
2) List all types of data, numeric, categorical, etc.
3) Perform EDA on data
Utilize both:
- Classic approach in EDA (Pandas, Numpy libraries)

- The TFDV (TensorFlow Data Validation) module with the powerful graphical statistics generated (apache beam library…)

Present dependencies and correlations among the various features in the data.
List the most variables (Feature Importance) that will affect the target label.

(B)
Build a **Deep Learning model** (based on **Tensorflow's classification modeling** approaches) that provides reliable and improved accuracy on classifying the correct class of a wine.

Typical architecture of a classification neural network
The word typical is on purpose.
Because the architecture of a classification neural network can widely vary depending on the problem you're working on.

However, there are some fundamentals all deep neural networks contain:
- An input layer.
- Some hidden layers.
- An output layer.

Much of the rest is up to the data analyst creating the model.
The following are some standard values you'll often use in your classification neural networks.

Hyperparameter	**Binary Classification**	**Multi-class Classification**
Input layer shape	Same as number of features (e.g. 5 for age, sex, height, weight, smoking status in heart disease prediction)	Same as binary classification
Hidden layer(s)	Problem specific, minimum = 1, maximum = unlimited	Same as binary classification
Neurons per hidden layer	Problem specific, generally 10 to 100	Same as binary classification
Output layer shape	1 (one class or the other)	1 per class (e.g. 3 for tree, person or animal photo)
Hidden activation	Usually ReLU (Rectified Linear Unit)	Same as binary classification
Output activation	Sigmoid	Softmax
Loss function	Cross entropy (tf.keras.losses.BinaryCrossentropy in TensorFlow)	Cross entropy (tf.keras.losses.CategoricalCrossentropy in TensorFlow)
Optimizer	SGD (Stochastics Gradient Descent), Adam	Same as binary classification

Perform **multi-class classification modeling analysis** on the ready-to-be-fed data.

In TensorFlow, there are typically 3 fundamental steps to creating and training a model.
- ✓ **Creating a model** - piece together the layers of a neural network yourself (using the functional or sequential API) or import a previously built model (known as transfer learning).
- ✓ **Compiling a model** - defining how a model's performance should be measured (loss/metrics) as well as defining how it should improve (optimizer).
- ✓ **Fitting a model** - letting the model try to find patterns in the data (how does X get to y).

Perform the following tasks:

- Evaluating and improving our classification model
- Split the dataset to 80%/20% ratio (training vs test)
- Evaluate your model on the test dataset.
- Plot the graphs of loss vs accuracy curves
- Print the Confusion Matrix

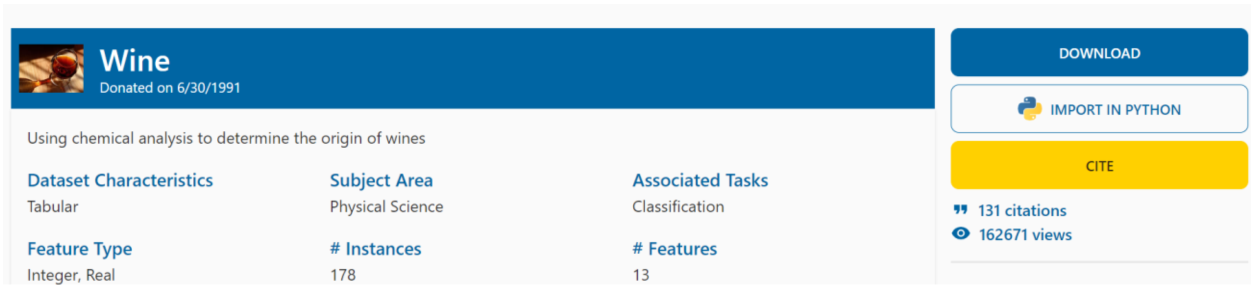
Tune your model on the following parameters:
The **activation parameter** - We used strings ("relu" & "sigmoid") instead of using library paths (tf.keras.activations.relu), in TensorFlow, they both offer the same functionality.

The **learning_rate** (also **lr**) parameter – Set a proper range of values for learning rate, usually from 0.1 down to 0.001 or less.
Important Note: You can think of the learning rate as how quickly a model learns. The higher the learning rate, the faster the model's capacity to learn, however, there's such a thing as a too high learning rate, where a model tries to learn too fast and doesn't learn anything. We'll see a trick to find the ideal learning rate soon.

The **number of epochs** - Remember a single epoch is the model trying to learn patterns in the data by looking at it once. Try several sizes: 500, 1000, 2000, etc...

Dataset / Content
Download the dataset from UC Irvine Machine Learning Repository:
https://archive.ics.uci.edu/dataset/109/wine

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

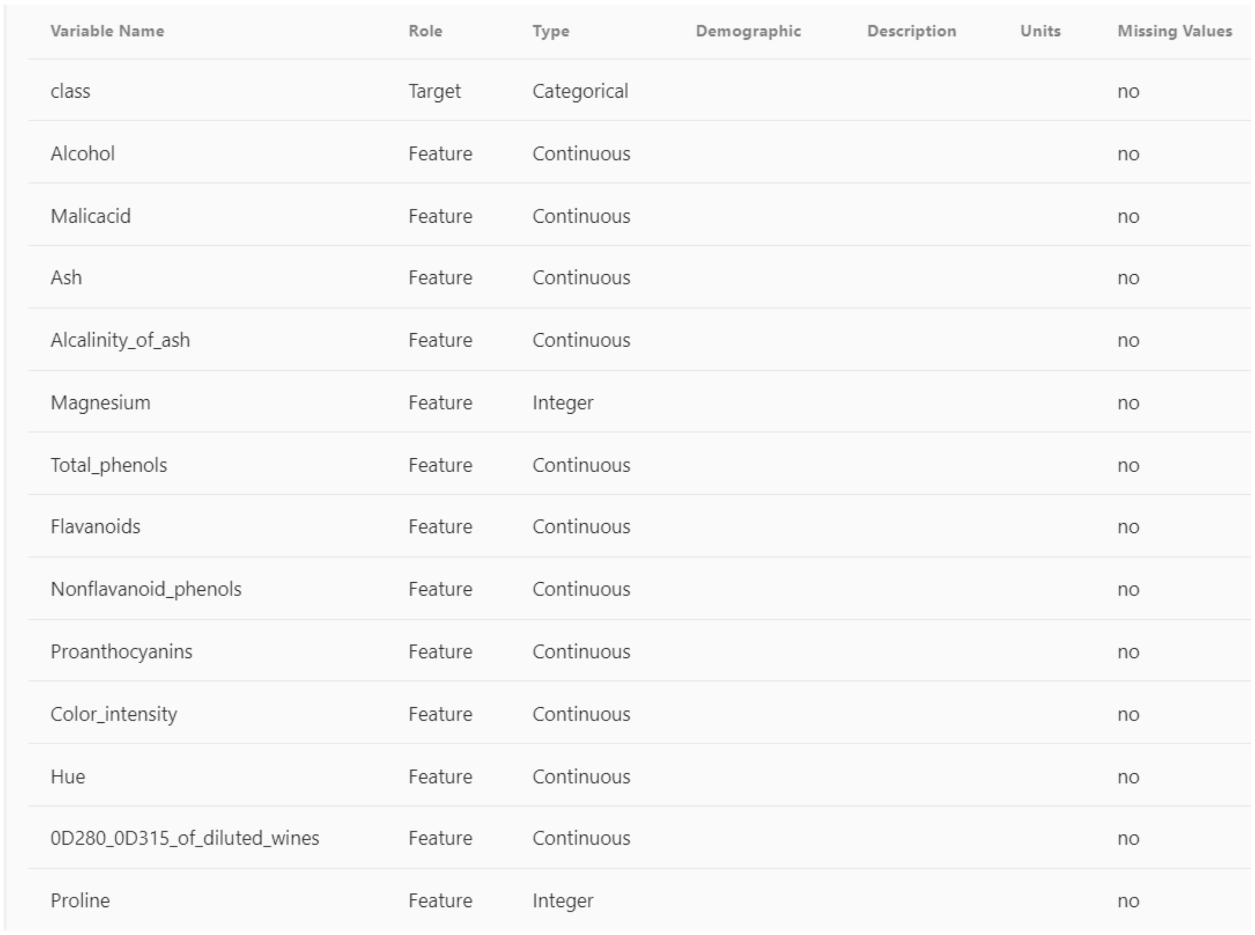

Wine
Donated on 6/30/1991

Using chemical analysis to determine the origin of wines

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Physical Science	Classification
Feature Type	**# Instances**	**# Features**
Integer, Real	178	13

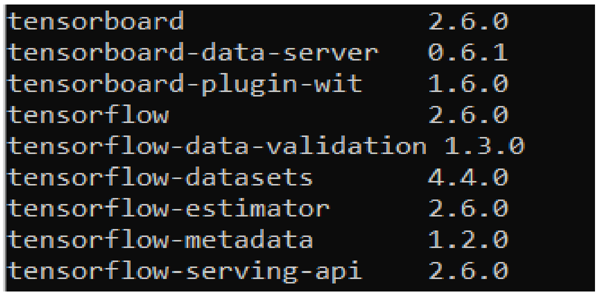
131 citations
162671 views

The columns are as follows: **13 features** and **1 target/label** with 3 possible values (1,2,3):



Variable Name	Role	Type	Demographic	Description	Units	Missing Values
class	Target	Categorical				no
Alcohol	Feature	Continuous				no
Malicacid	Feature	Continuous				no
Ash	Feature	Continuous				no
Alcalinity_of_ash	Feature	Continuous				no
Magnesium	Feature	Integer				no
Total_phenols	Feature	Continuous				no
Flavanoids	Feature	Continuous				no
Nonflavanoid_phenols	Feature	Continuous				no
Proanthocyanins	Feature	Continuous				no
Color_intensity	Feature	Continuous				no
Hue	Feature	Continuous				no
0D280_0D315_of_diluted_wines	Feature	Continuous				no
Proline	Feature	Integer				no

(**) Highly recommend to have installed the whole gamma of TensorFlow's module.
Here is a 'base' list of them:



```
tensorboard                 2.6.0
tensorboard-data-server     0.6.1
tensorboard-plugin-wit      1.6.0
tensorflow                  2.6.0
tensorflow-data-validation 1.3.0
tensorflow-datasets         4.4.0
tensorflow-estimator        2.6.0
tensorflow-metadata         1.2.0
tensorflow-serving-api      2.6.0
```