# ▤ TASK 2

## Disease Classification

You are in the group 👥 Chewbhan consisting of 👤 alefevre (alefevre@student.ethz.ch (mailto://[u'alefevre@student.ethz.ch'])), 👤 mwatine (mwatine@student.ethz.ch (mailto://[u'mwatine@student.ethz.ch'])) and 👤 nescarfail (nescarfail@student.ethz.ch (mailto://[u'nescarfail@student.ethz.ch'])).

## 📖 1. READ THE TASK DESCRIPTION

## 🖥 2. SUBMIT SOLUTIONS

## ✉ 3. HAND IN FINAL SOLUTION

# 📖 1. TASK DESCRIPTION

## TASK 2: DISEASE CLASSIFICATION FROM IMAGE FEATURES

This task is primarily concerned with multi-class classification where you have 3 classes. However, we have changed the original image features in several ways. You will need to deal with class imbalance; in the training set, there are 600 examples from class 0 and 2 but 3600 examples from class 1. Test set has the same class imbalance as the training set.

## DATA DESCRIPTION

Download Data (/static/task2.zip)

The data for this task contains the following files:

- **X_train.csv** - the training features
- **y_train.csv** - the training targets
- **X_test.csv** - the test features (you need to make predictions for these samples)
- **sample.csv** - a sample submission file in the correct format

Each row in X_train.csv is one sample indexed by an id, so the first column contains the id. In addition to the id column, each sample has 1000 features:

```
id,x0,x1,...,x999
0,-1.09,0.91,...,2.44
...
```

The test set file (X_test.csv) has the same structure:

```
id,x0,x1,...,x999
0,0.39,-1.26,...,0.59
...
```

The training targets (image classes) are contained in y_train.csv:

```
id,y
0,1
1,0
2,1
...
```

For your convenience, we further provide a sample submission file (sample.csv):

```
id,y
0,0
1,0
2,0
...
```

Note that, for each prediction you need to include the id of the sample as in X_test.csv.

## SUBMISSION FORMAT

**For every data instance in the test set**, submission files should contain two columns: *id* and *y* where *y* should be the class labels (0, 1 or 2).

The file should contain a header and have the following format:

```
id,y
0,1
1,2
2,0
...
```

Please keep in mind that, as a group, you have a limited number of submissions as stated on the submissions page.

## EVALUATION

The evaluation metric for this task is the **Balanced Multi Class Accuracy (BMAC)** which ranges from 0 to 1.
In other words, it is the average recall of each class. Mathematically, the formula is:

$$BMAC = \frac{1}{C} \sum_{c=1}^{C} TPR_c$$

where $TPR_c$ is the true positive rate (sensitivity, recall) for class c.

How to compute it in Python:

```
from sklearn.metrics import balanced_accuracy_score
BMAC = balanced_accuracy_score(y_true, y_pred)
```

> ⚠ Make sure that you properly hand in the task, otherwise you may obtain zero points for this task.

## FREQUENTLY ASKED QUESTIONS

### WHICH PROGRAMMING LANGUAGE AM I SUPPOSED TO USE? WHAT TOOLS AM I ALLOWED TO USE?

You are free to choose any programming language and use any software library.

### CAN YOU HELP ME SOLVE THE TASK? CAN YOU GIVE ME A HINT?

As the tasks are a graded part of the class, **we cannot help you solve them**. However, we will try to address general aspects during the project tutorials. Moreover, feel free to ask general questions about the course material during or after the exercise sessions.

### CAN YOU GIVE ME A DEADLINE EXTENSION?

> ⚠ We can not grant deadline extensions, except in extraordinary cases (e.g. military service). However, we will require official confirmation of your problem (e.g. certificate of illness).

### CAN I POST ON PIAZZA AS SOON AS I HAVE A QUESTION?

This is highly discouraged. Instead,

- Read the details of the task thoroughly.
- Review the frequently asked questions.
- If there is another team that solved the task, try again.
- Discuss with your team mates.

If you still consider that you should contact the TAs, you can post a **private** question on Piazza. Remember that collaboration with other teams beyond (general discussions) is prohibited.

### WHEN WILL I RECEIVE THE PRIVATE SCORES? AND THE PROJECT GRADES?

We will publish the private scores before the exam the latest.