# HW#2 – Machine Learning in Healthcare 336546

This assignment relates to the detection of Type 1 Diabetes (T1D) from a simple yes/no questionnaire, asking patients about their medical history. Your goal is to predict if a patient has T1D by applying ML algorithms on this dataset.

## Type 1 Diabetes (also known as juvenile diabetes)

T1D is a chronic condition resulting from a lack of insulin in the body. The disease typically presents in early childhood or adolescence. Up to 0.33% of the global population suffers from T1D, making it a world-wide and wide-spread issue. There is no cure and the current treatment is to control blood glucose levels through glucose monitoring, insulin injections, diet, and lifestyle modifications to prevent complications.

The exact cause of T1D is a mystery; however, there are few possible causes such as genetics, autoimmune dysfunction or environmental factors such as some kind of viruses.

Credit: https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/symptoms-causes/syc-20353011

**Assignment**

This assignment has no prebuilt notebook. You are required to build and present an appropriate notebook to show your experiments and results. Please provide all answers within the notebook (in a markdown cell), labeled carefully based on the question number. In this assignment, you will do the following:

- Explore the data provided.
- Implement linear and non-linear classifiers.
- Model optimization with k-fold cross validation
- Evaluate your model performances with appropriate metrics.
- Present a 2d visualization of multi-featured data.
- Use feature selection tools.

Use the provided HW2 environment and any additional packages you need for this assignment.

## Theory Questions (28%)

1) To evaluate how well our model performs at T1D classification, we need to have evaluation metrics that measures of its performances/accuracy. Which evaluation metric is more important to us: model accuracy or model performance? Give a simple example that illustrates your claim.

2) T1D is often associated with other comorbidities such as a heart attack. You are asked to design a ML algorithm to predict which patients are going to suffer a heart attack. Relevant patient features for the algorithm may include blood pressure (BP), body-mass index (BMI), age (A), level of physical activity (P), and income (I). You should choose between two classifiers: the first uses only BP and BMI features and the other one uses all of the features available to you. Explain the pros and cons of each choice.

3) A histologist wants to use machine learning to tell the difference between pancreas biopsies that show signs of T1D and those that do not. She has already come up with dozens of measurements to take, such as color, size, uniformity and cell-count, but she isn’t sure which model to use. The biopsies are really similar, and it is difficult to distinguish them from the human eye, or by just looking at the features. Which of the following is better: logistic regression, linear SVM or nonlinear SVM? Explain your answer.

4) What are the differences between LR and linear SVM and what is the difference in the effect/concept of their hyper-parameters tuning?

## Coding Assignment (72%)

The data for this exercise can be found in the attached file named HW2_data.csv. There are 565 patients in the database. The nurse who collected the data said that not all patients answered all the questions.

1) Load the data. Explain any preprocessing. (5%)

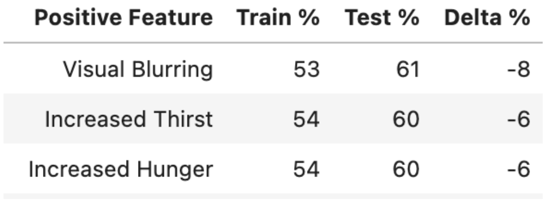
2) Perform a test-train split of 20% test. (5%)

3) Provide a detailed visualization and exploration of the data. (10%)

   You should at least include:

   a. An analysis to show that the distribution of the features is similar between test and train. See table 1 below.
      i. What issues could an imbalance of features between train and test cause?
      ii. How could you solve the issue?
   b. Plots to show the relationship between feature and label. See Figure 1 below.
   c. Additional plots that make sense given the mostly binary nature of this dataset.
   d. State any insights you have
      i. Was there anything unexpected?
      ii. Are there any features that you feel will be particularly important to your model? Explain why.

*Table 1: Table showing distribution between each feature label in Train and Test Sets*



| Positive Feature | Train % | Test % | Delta % |
|---|---|---|---|
| Visual Blurring | 53 | 61 | -8 |
| Increased Thirst | 54 | 60 | -6 |
| Increased Hunger | 54 | 60 | -6 |

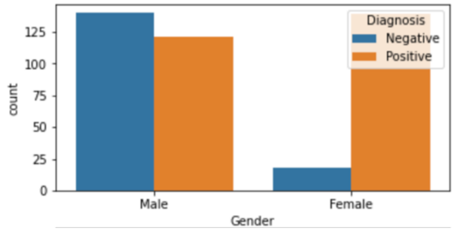


*Figure 1: Plot showing the frequency of Increased Urination according to Diagnosis*

4) **Encode** all your data as one hot vectors. (5%)

5) Choose, build and optimize Machine Learning Models: (20%)
   a. Use 5k cross fold validation and **tune** the models to achieve the highest test AUC:
      i. Train one or more linear model on your training set
      ii. Train one or more non-linear models on your training set
   b. Report the appropriate evaluation metrics of the train and test sets (AUC, F1, LOSS, ACC).
   c. What performs best on this dataset? Linear or non-linear models?

6) Feature Selection (10%)

   a. As seen previously, a Random Forest Network can be used to explore feature importance. Train a Random Forest on your data.

      i. What are the 2 most important features according to the random forest.

      ii. Does this match up exactly with the feature exploration you did?

Note: Question 7 should only be completed after your lecture on dimensionality reduction

7) Data Separability Visualization: (20%)

   a. Perform dimensionality reduction on the dataset so that you can **plot your data in a 2d plot** (show samples with positive and negative labels in different colors).

   b. How separable is your data when reduced to just two features?

   c. Train the same models above on the dimensionality-reduced training set.

   d. Train the same models on the best two features from section 6.

   e. What performs better? 2 features of the reduced dimensionality.