**Team members:** Mael Illien, Salma Elshahawy

## Introduction

The ability to obtain a rapid understanding of the context of a patient’s overall health can be crucial to medical outcomes. Patients go to hospitals for a variety of reasons. These reasons range from benign to severe and the state of patients can also range from alert to unconscious. Gathering information that is important for treatment requires processing of patient data but the lack of verified medical histories and the structural inefficiencies of obtaining medical records that take days to transfer pose a problem for medical staff who would benefit from quickly knowing the presence of certain chronic conditions such as heart disease, injuries, or diabetes in a patient. Additionally, patients can also be uncooperative or untruthful and withhold information. For these reasons, the ability to detect chronic conditions from basic data can be of great value to make informed clinical decisions.

## Background

Diabetes Mellitus, or simply diabetes is one of these chronic conditions that is important for medical practitioners to be aware of when treating patients. Detecting the presence of diabetes from patient data is a supervised machine learning problem of binary classification and a number of tools such as discriminant analysis, logistic regression and naive Bayes can be utilized for detection of diabetes. Analysis of the dataset (specify dataset) can help discover insights useful for model building, inference and prediction.

## Why it’s interesting

This is an interesting problem because it is not limited to diabetes. Any number of chronic health conditions can be detected using the same methodology. It is also useful because it may be used to detect conditions that patients may not even know they had. This is a step in the direction of providing better patient specific healthcare.

## Method and Literature Review

Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease - This paper evaluates whether miRNA expression profiling in conjunction with machine learning classification techniques is a suitable non-invasive test to diagnose inflammatory bowel disease (IBD), in particular Crohn's disease (CD) and ulcerative colitis (UC). The ML methods evaluated are penalized SVM models, namely LASSO SVM, elastic net SVM, SCAD SVM and elastic SCAD SVM. To evaluate the validity of the feature selection employed by the penalized SVMs, two ensemble random forests models were built for each classification problem.

Prediction of Diabetes using Classification Algorithms - This paper uses three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes to detect diabetes at an early stage using data from the Pima Indians Diabetes Database (PIDD) sourced from the UCI machine learning repository

Classification and prediction of diabetes disease using machine learning paradigm - This paper uses Logistic Regression to identify the risk factors for diabetes and four classifiers, namely Naïve Bayes, Decision Trees, Adaboost, and Random Forest to predict the occurence of diabetes in patients.

Diabetes is normally diagnosed using either:

- Oral Glucose Tolerance Test OGTT: measures your body's response to sugar (glucose). This requires an overnight fast, a post fast blood sugar reading as a baseline followed by the ingestion of a glucose solution and supplementary blood sugar readings. A normal blood glucose level is lower than 140 mg/dL (7.8 mmol/L). A blood glucose level between 140 and 199 mg/dL (7.8 and 11 mmol/L) is considered impaired glucose tolerance, or prediabetes. A blood glucose level of 200 mg/dL (11.1 mmol/L) or higher may indicate diabetes.
- A1C Test: also called the glycated hemoglobin, glycosylated hemoglobin, hemoglobin A1C or HbA1c test. An A1C test result reflects the average blood sugar level for the past two to three months. The A1C test measures the percentage of hemoglobin proteins in the blood are coated with sugar (glycated). The higher the A1C level is, the poorer a patient's blood sugar control is and the higher the risk of diabetes complications. A test value below 5.7% is normal. 5.7% to 6.4% is diagnosed as prediabetes. Readings of 6.5% or higher on two separate tests indicates diabetes.

## Hypothesis

The presence of Diabetes Mellitus can be detected from patient data collected within the first 24 hrs in an Intensive Care Unit. A parsimonious binary classification model can be built using variable selection and regularization techniques. The classification threshold can be adjusted to maximize sensitivity (true positive rate) of detection.

## Conclusion

The proposed solution solves the problem by providing medical staff with another tool for quick diagnosis without the need to potentially costly and slow tests.

## Data Source

https://www.kaggle.com/c/widsdatathon2021/data