# OVERVIEW

Diabetes mellitus is a chronic disease where the concentration of glucose in the blood of the patients is at an unusually high level (WHO, 2014). This disease could be caused by the failure of some cells in the body to respond to insulin, or the inability of the pancreas to produce insulin; a protein that stimulates blood glucose reduction (WHO, 2013). In the early stages of the condition, symptoms include frequent urination whilst at the end stage, is characterized by blindness, idiopathic foot ulcers, stroke and eventually death.

Approximately **8%** of the world population (adults only) suffer from diabetes IDF (2019). According to the CDC (2020), diabetes ranks 7th among the diseases that cause the most deaths in the worldwide. Though several efforts have been directed towards the reduction of the incidence of diabetes, cases continue to rise (IDF, 2019).

The risk of complications from diabetes can be prevented when measures aimed at controlling the blood glucose levels are applied at an early stage of the condition thus, making early detection and treatment of the condition very important. Also, diabetes tests in regions around the world that do not have ready access to healthcare takes days and sometimes months.

Therefore, it is very important to have a machine learning model, that can predict the diabetic status faster than the traditional methods. This would ensure that interventions to reduce the risk of complications with regards to diabetes, are reduced.

This problem will be a **binary classification task** and characteristics of patients such as blood glucose levels, skin thickness and age would be considered depending on the strength of correlation with diabetes outcome and would be used to predict whether a patient is diabetic or not.

To solve the problem, a model which couid be a KNN or ensemble model would be created that can classify an individual’s diabetes status based on some measured metrics such as blood glucose concentration and Body Mass Index (BMI). The dataset used in this project is a Pima Indian Diabetes dataset downloaded from Kaggle. Each row represents a patient and each column contains certain measured characteristic of the patient; Blood glucose Concentration, BMI, Skin Thickness, Insulin, Diabetes Pedigree Function and Outcome, which indicates whether the patient is diabetic or not. Portion of this dataset will be used to train a machine learning algorithm and the remaining portion of that dataset will be used for checking the out-sample performance of the algorithm.

The performance of each model will be checked by using classification accuracy, precision and F1 score. A confusion matrix, as well as a receiver operating characteristic (roc) curve would be used to determine the specificity and sensitivity of the model and the area under curve (auc) of our model respectively. This would provide us with a true picture of the performance of our model.