**Data Science Project**

**Early Stage Diabetes Risk Prediction**

Diabetes is one of the fastest growing chronic life-threatening diseases that have affected 422 million people worldwide according to the World Health Organization (WHO), in 2018. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. Around 50% of all people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase. The early diagnosis of diabetes is only possible by proper assessment of both common and less common signs and symptoms, which could be found in different phases from disease initiation up to diagnosis

**How to solve the problem**

A machine learning model would be developed to predict whether an individual is at risk of developing diabetes based on some signs and symptoms. This will help in the early detection and diagnosis of diabetes.

**Dataset**

The dataset was retrieved from the UCI Machine Learning Repository. It contains the signs and symptoms data of newly diabetic or would be diabetic patients. This has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

The dataset contains **520** rows and **17** attributes.

**Attribute Information:**

|  |  |
| --- | --- |
| Age | 20-65 |
| Sex | 1. Male, 2. Female |
| Polyuria | 1. Yes, 2. No. |
| Polydipsia | 1. Yes, 2. No. |
| sudden weight loss | 1. Yes, 2. No. |
| weakness | 1. Yes, 2. No. |
| Polyphagia | 1.Yes, 2. No. |
| Genital thrush | 1. Yes, 2. No. |
| visual blurring | 1. Yes, 2. No. |
| Itching | 1. Yes, 2. No. |
| Irritability | 1.Yes, 2.No. |
| delayed healing | 1.Yes, 2.No. |
| partial paresis | 1.Yes, 2.No. |
| muscle stiffness | 1.Yes, 2.No. |
| Alopecia | 1.Yes, 2.No. |
| Obesity | 1.Yes, 2.No. |
| Class | 1.Positive, 2.Negative. |

**Why use machine learning**

The application of ML in healthcare has received many attentions and produced promising outcomes in the identification and diagnosis of diseases. Using ML is the best solution to this problem as the algorithms are quick and effective in detecting trends and pattern in the already existing data available.

**Why this data set**

Though there are a lot of ML applications on diabetes datasets, most features in the dataset used were based on biological measurements or laboratory test conducted on the patient. The process of getting such data can become cumbersome for an individual who wants to use the model to predict whether he or she is likely to develop diabetes.

This dataset was derived directly through questionnaires from patients in Sylhet, Bangladesh. Since Bangladesh is classified as a third world country like Ghana, they are likely to share similar demographic characteristics, economic and social status. This could be argued that building a model on such data could be relevant in the Ghanaian setting.

Also, since the data used requires only the response of patients, it will be easy for anyone to use this current model when it is developed and deployed.

**Implications**

Having a technology to accurately predict if an individual will have diabetes will help in the early diagnosis of the condition in other to avoid further progression to advanced stage or possibly death.