Project Proposal

[1] **What is the topic/title of your research project?**

Predictive Analysis for Diabetes Detection among Patients.

[2] **What research questions does your project initially aim to address? Formulate your research questions rigorously.**

When the blood glucose, commonly known as blood sugar, becomes too high, you develop diabetes. Your primary energy source is blood glucose, which is obtained out from food we eat. The pancreas produces the hormone insulin, which facilitates the entry of food-derived glucose into your cells for energy production. Your body occasionally produces insufficient or no insulin, or it uses insulin poorly. After that, glucose remains in your circulation and does not enter your cells. Diabetes is a disease that is prevalent, has a plethora of data available, and comes with significant problems. There is a requirement for a more effective and precise method of disease diagnosis.

The project aims at detecting if the patient being studied has diabetes or not. If the patient does not have diabetes then it also suggests that the person is prone to developing diabetes in the long run. This is achieved by means of various physiological and pathological experiments from samples collected from the patients.

[3] **What dataset(s) have you considered to use for your project? Be specific and mention your source(s).**

The information was gathered from Iraqi society, as well as information from the lab at Medical City Hospital and (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital). To create the diabetes dataset, data from patient files were obtained, extracted, and entered the database. Medical information and laboratory analysis make up the data.

* **Name** - Diabetes Dataset
* **Source**- <https://data.mendeley.com/datasets/wj9rwkp9c2/1>
* **Format** - Comma Separated Value(.csv)

[4] **What is the total size of your dataset? How many instances (i.e. rows) does your dataset possess?**

The ‘Diabetes Dataset’ which has data for the pathological examinations of patients sourced from <https://data.mendeley.com/datasets/wj9rwkp9c2/1> has the following size or dimensions:

* **Size** - 49 kB
* **Columns** - 14
* **Rows** - 1000

[5] **What are the potential features for your research problem? How many potential features do you have?**

The ‘Diabetes Dataset’ has the following set of features or predictor variables which would play a role in the detection of whether a patient would have diabetes, has diabetes or does not have diabetes. The features are as follows:

* ID - Id
* No\_Pation - No. of Patient
* Gender - Gender of Patient
* AGE - Age of patient
* Urea - Urea level
* Cr- Creatinine ratio(Cr)
* HbA1c- Glycated Haemoglobein
* Chol- Cholesterol (Chol)
* TG- Triglycerides(TG)
* HDL-High density lipoprotein
* LDL- Low density lipoprotein
* VLDL- Very Low density lipoprotein
* BMI- Body Mass Index (BMI)

[6] **What are your target variables? Specify the nature of your target variables (i.e. continuous or categorical).**

In our domain of study, the target variable which is the ‘Class’ column in the diabetes dataset is a categorical variable. The class variable is mainly composed of 3 different values according to our dataset. The classes can be described as follows:

* **Y** - Diabetic Patient
* **N** – Non-Diabetic Patient
* **P** - Predict Diabetic Patient