# **CHAPTER 3**

# **RESEARCH METHADOLOGY**

## **3.1 INTRODUCTION**

This chapter explains in details the methods and materials to be used for the success of the research.

## **3.2 RESEARCH DESIGN**

The research will be done in the following steps in order to achieve the stated objectives:

* Date Collection
* Data Processing/ Analysis
* Feature Selection
* Classification model
* Implementation/ Development of Predictive Model
* Performance Evaluation

## **3.3 DATA COLLECTION**

……..Definition of data collection and explanation of the how we collected data from Nizamiye Hospital,……

Then:

Table 3.1 Features Information

|  |  |
| --- | --- |
| Header | Feature Information |
| Patient name | Names |
| Fasting Blood Glucose | Glocose level in a person’s blood when the person is hungry |
| Hct | definition |
| Hgb | definition |
| Mch | definition |
| Mchc | ……… |
| Mcv | ……… |
| Pdw | ……. |
| Plt | ……………. |
| Rbc | ………….. |
| Rdw-Cv | …………… |
| Rdw-Sd | ………….. |
| Wbc  Result | ………………  There is diabet or not (positive or negative) |

## **3.4 DATA PROCESSING/ ANALYSIS**

Definition of Data Processing ………….

To perform the cleaning of the data, four modules in **Python** will be used, they include Pandas, NumPy, Matplotlib, and Seaborn. (in a research on a sickness they have used these models)Then the following steps will be taken:

1. The first step is to understand the data, and the following analysis will be done to the parts of the data that are numerical:

* Indicating the number of entries
* Calculating the mean of the values
* Calculating standard deviation of each features value.
* Identifying the min and the max value in the data and the corresponding points in between.

1. A group charts showing relationship between;

* Fasting blood glucose and diabetes.
* Hct and diabetes.
* Hgt and diabetes.
* Mch and diabetes.
* Mchc and diabetes.
* Mcv and diabetes.
* Pdw and diabetes.
* Plt and diabetes.
* Rdw and diabetes.
* Rdw-Cv and diabetes.
* Rdw-Sd and diabetes.
* Wbc and diabetes.

All of this charts will be drawn in other to understand the relationship between the features and the target features (Result: positive or negative, there is diabetes or not )

1. The next step will be to perform correlation analysis, in order to identify the blood component features that are key.

A correlation plot will be plotted to find the correlation between the many features of the data.

1. Then finally encoding the Result of Diabetes as 0 (negative) and 1 has (positive)

## **3.5 FEATURE SELECTION**

The blood components features to be used for the building of the model will be decided in the data cleaning stage, after correlation analysis and and dispersion matrix (no idea what they are for NOW but I ll learn) is done to further analyze the importance’s of each feature from the data. The selected features among blood components will be indicated below :

* ……
* …….
* …….
* ……..
* ………
* …….
* …….
* ……..
* …….

## **3.6 CLASSIFICATION METHOD**

A brief explanation about data mining and importance of data mining in medical area.

A brief explanation of Random Forest and importance of this method and why we choose this method ……..

### 3.6.1 How Random Forest Works?

…………………………

## **3.7** **MODEL PERFORMANCE EVALUATION**

Every Model built has to be tested, in order to ascertain its accuracy. The proposed model will be evaluated/tested with the test data to know its predictive skill .

### 3.7.1 Confusion Matrix

Definition , how and why it is used, explanation of the components of confusion matrix

### 3.7.2 Accuracy

Definition, how it is found.

## **3.8 REQUIREMENT ANALYSIS**

### 3.8.1 Tools and Platforms

Why we will use python.

### 3.8.2 Seaborn

Definition , usage

### 3.8.2 Pandas

Definition , usage

**3.8.3** **NumPy**

Definition , usage

**3.8.4** **Matplotlib**

Definition , usage

3.8.5. anyother