



## Programming Fundamentals

### Assignment 02: HCS - Process data

#### 1 Introduction

With the development of technologies, we now have various ways to access health-care systems. One uses a smart watch can check their heart rate anytime, including the history of his movements. Some advanced IoT devices can even track your blood sugar level and send those information to provide a fast diagnosis. The health-care systems will soon become a part of our life in this digital world. In this assignment, the student has to build a very simple program that can simulate some features of a health-care system.

In this assignment, student must build a system that can perform analysis on the medical data. The data in this assignment is “Pima Indians Diabetes Database”<sup>1</sup> from Kaggle. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Student is given a data file named “diabetes.csv” in CSV format (Comma-separated values):

```
1 Pregnancies , Glucose , BloodPressure , SkinThickness , Insulin , BMI, DiabetesPedigreeFunction , Age ,  
   Outcome  
2 2,138,62,35,0,33.6,0.127,47,1  
3 0,84,82,31,125,38.2,0.233,23,0  
4 0,145,0,0,0,44.2,0.63,31,1  
5 0,135,68,42,250,42.3,0.365,24,1  
6 1,139,62,41,480,40.7,0.536,21,0  
7 0,173,78,32,265,46.5,1.159,58,0  
8 4,99,72,17,0,25.6,0.294,28,0  
9 8,194,80,0,0,26.1,0.551,67,0  
10 2,83,65,28,66,36.8,0.629,24,0  
11 2,89,90,30,0,33.5,0.292,42,0  
12 4,99,68,38,0,32.8,0.145,33,0  
13 4,125,70,18,122,28.9,1.144,45,1  
14 3,80,0,0,0,0,0.174,22,0  
15 6,166,74,0,0,26.6,0.304,66,0  
16 5,110,68,0,0,26,0.292,30,0  
17 2,81,72,15,76,30.1,0.547,25,0  
18 7,195,70,33,145,25.1,0.163,55,1  
19 6,154,74,32,193,29.3,0.839,39,0  
20 2,117,90,19,71,25.2,0.313,21,0  
21 3,84,72,32,0,37.2,0.267,28,0  
22 6,0,68,41,0,39,0.727,41,1  
23 7,94,64,25,79,33.3,0.738,41,0  
24 3,96,78,39,0,37.3,0.238,40,0  
25 10,75,82,0,0,33.3,0.263,38,0  
26 0,180,90,26,90,36.5,0.314,35,1
```

Student will write the program to process requests on this dataset. The details of requests are given in the next section.

The student should perform a deep analysis of problem requirements before start implementing. In addition, the student is encourage to follow the problem solving process in order to do well in the assignment. In this assignment, we focus on using control structures, including loops, functions, etc. Note that, NO class implementation is allowed in this assignment.

<sup>1</sup>[www.kaggle.com/alhassanabdelgilil/predicting-the-diabetes/data](http://www.kaggle.com/alhassanabdelgilil/predicting-the-diabetes/data)

## 2 Requirements

The meaning of each column is given as follows:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight (kg) / height ( $m^2$ ))
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

In this assignment, we will receive a sequence of request from the client. These requests are represented in form of input text as follows:

Request	Output	Description
CR	number	Count the number of records in the dataset.
DI <info> <col>	number	Calculate information of a column in the dataset. <info> is the information we need to calculate: Mean, Standard-Deviation, Min, Max <col> is the data we need to calculate: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigree-Function, Age
HI <col> <min> <max> <step>	integer array	Calculate the histogram of one particular column in the dataset.  <col> is the data we need to calculate. <min> the minimum value of the column. <max> the maximum value of the column. <step> the size of the bins. Example: for a request HI Glucose 0 100 10 The output will be an integer array where $i^{th}$ number is the number of records that has Glucose in range $[i * 10, i(+1) * 10)$
FR	integer array	A list of indexes of records indicating Outcome = 1.
FR <col> <min> <max>	integer array	A list of indexes of records indicating Outcome = 1 and has <col> in the range [<min>, <max>].

The requests are given in form of strings, and the location of the output is given in an input address. Student must implement the function to load data from the file into the defined struct.

## 3 Implementation

An initial code package is provided with some source file and configuration files. The main.cpp file contain the main function of the application and should not be modified since it will be overwritten during grading process. Two files core.cpp and core.h will be used for implementing features for this program.

In this first assignment, student can use provided functions, global variables. However, student can also write other functions if needed. Remember, file `core.h` is where you will write prototypes and `core.cpp` is the place to put your implementation.

## 4 Regulation

### 4.1 Evaluation

The program output will be compared with the expected output from the solution. A testcase is passed if everything in the program output is matched to the solution. The testcase will fail if the program runs too slow (timeout error). This issue should not occur in this assignment but the student must be aware.

### 4.2 Submission

Student should follow the submission instruction on the course site (e-learning site). One file `src.zip` contains `core.cpp` and `core.h` must be submitted during the given timeline. The OS system that builds and runs code is Linux. Please do not use any special functions that do not exist on Linux, otherwise your code cannot be built. Student can build their code on Linux before submit to ensure that the code works.

The deadline for implementing this assignment is 07-05-2019. The grading system will open soon.

### 4.3 Rules

Student must perform this assignment themselves. Copying code is strictly prohibited and be considered as ethic code violation. In such case, the course result will be zero regardless the upcoming assignments. Please protect your code carefully. Student can help their friends by giving advice, not code.