|  |  |  |
| --- | --- | --- |
| A picture containing text, clipart, soup, dish  Description automatically generated  **Credit Hours System** |  | A picture containing text, clipart  Description automatically generated  **Cairo University**  **Faculty of Engineering** |

Patient Hospital Stay Prediction

Phase 1

Ahmed Habib - 1170323

Ahmed Osama - 1180075

Yumna Hamdy

Rawda Rumaih

Youssef

# Introduction

During Covid-19 pandemic, healthcare systems crashed in many countries due to the overload on hospitals and not being able to predict how long a patient would stay to better estimate the capacity of the hospital and introduced the problem for the patients which is which hospital should they go to.

This raised the need for data science and machine learning to hopefully find a solution for such a problem that’s even before Covid-19 many hospitals suffered from.

# Data

For this problem we collected data from Analytics Vidhya Hackathon about Healthcare Analytics a hackathon that was held back in late 2020 with the same problem statement as we proposed.

## Data description:

The data has information about the patients and the hospitals and the duration of stay, and we have 318,438 training examples for 18 features.

### Features:

1 – case\_id: A unique id for the case in the hospital

2- Hospital\_Code: A unique code for hospital

3- City\_Code\_Hospital: Code for the City in which the hospital is located

4- Hospital\_region\_code: Region Code of the Hospital

5- Available Extra Rooms in Hospital: Number of Extra rooms available in the Hospital

6- Department: Department overlooking the case

7- Ward\_Type; Code for the Ward type

8- Ward\_Facility\_Code: Code for the Ward Facility

9- Bed Grade: Condition of Bed in the Ward

10- patientid: Unique Patient Id

11- City\_Code\_Patient: City Code for the patient

12- Type of Admission: Admission Type registered by the Hospital

13- Severity of Illness: Severity of the illness recorded at the time of admission

14- Visitors with Patient: Number of Visitors with the patient

15- Age: Age of the patient

16- Admission\_Deposit: Deposit at the Admission Time

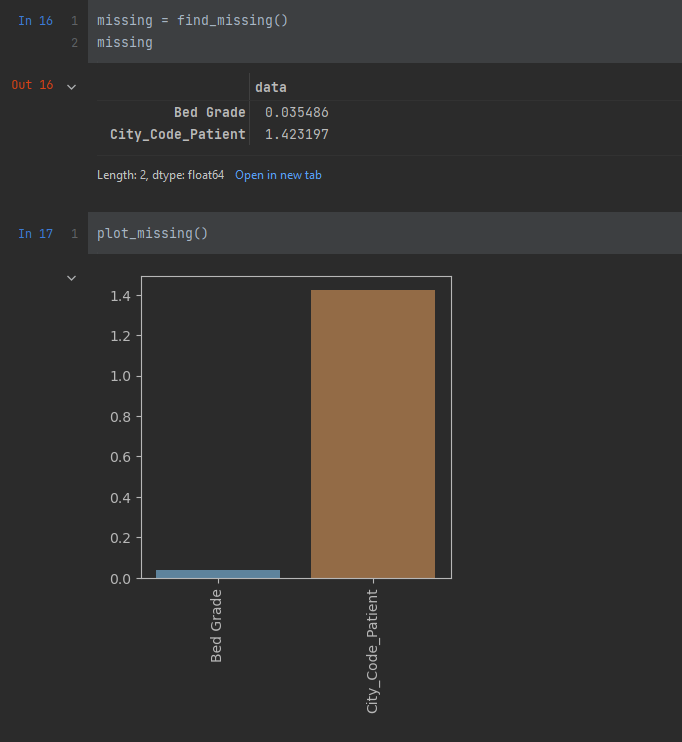
17- Stay: Stay Days by the patient, split in categories with values of:

0-1011-2021-3031-4041-5051-6061-7071-8081-9091-100More than 100 Days

## Our Proposed Solution:

Create a machine learning classification model that can classify the length of stay for the patient based on the other features collected in the dataset we have, and following the data science pipeline:

### Data Wrangling:

For the data wrangling phase we needed to perform an exploratory data analysis on our dataset to get a sense of the data and the features we have, so for the missing data, the missing data wasn’t really missing that much, 

(values in the above image are percentages).

Only for 2 features and very few is missing, so for the bed grade I went on and filled it with the mode of the bed grade values which was 2, since it’s categorical from 2 to 5.

But for the city code patients, I though of filling it with the patient city code but I barely found any correlation between the two so it wasn’t really reasonable to do so, and then I went on and dropped these missing values.

For Outliers:

We only had 3 numerical features to deal with outliers with which are:

Available extra rooms in hospital – visitors with patient – admission deposit