**Project Proposal**

**This document will talk about the proposal of my project for SDAIA Bootcamp, I will talk about the idea and more details below**

* **Domain Background:**

These days, a lot of people are facing issues that are related to cardiovascular or heart diseases that cause a really serious health conditions to a lot of people, and sometime they aren’t enough time so we can detect if the patient might face heart disease or not.

* **Problem Statement & Question:**

Like I said in the background section, heart disease are really serious health conditions that might cause death, a lot of people die sadly due to heart diseases and some times there are not enough time to diagnoses many people at once If there are not enough doctors or no enough time to detect or in one case or we can detect if the patient will face a cardiovascular disease or not

So our question or need is

Can we detect and predict if the patient might have cardiovascular diseases or not?

* **Datasets Description:**

The shape of data we have is

70,000 rows with 13 columns or features, where each row describe one patient

There are 3 types of Inputs we have in the dataset:

1. Objective: Factual Information
2. Examination: results of a medical test or exam
3. Subjective: Information was given by the patient him self

And the features or columns we have are 12 and they are as following:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap\_hi | int |
6. Diastolic blood pressure | Examination Feature | ap\_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. **Presence or absence of cardiovascular disease | Target Variable | cardio | binary | (Our Target)**

* **Solution Statement:**

One of the solutions are to detect the pattern of cardiovascular disease based on historical data of patients who visited hospital and have been diagnosed with heart diseases, so we might find two solutions, one is to detect it early before it happens, second to detect if the patient might have it but he doesn’t know based on his data

* **Benchmark model:**

One of the models can be found in this link on a Kaggle project: <https://www.kaggle.com/prathamsinghce/heart-failure-predictor>

Where the model predicted 84% of accuracy and we will be comparing to it

* **Evaluation metrics:**

Based on the goal or solution statement, we can focus on two things but mainly is the Recall, since we want the model to correctly classify and understand the pattern of heart diseases, also we can check the accuracy of the final model.

* **Project Design & Tools:**

First there will be an Exploratory data analysis or known as EDA, so we can understand the data more and find insights and analysis, also clean the data more if needed, and we considering logistic regression or K- nearest neighbor models, since we have a binary classification model.

For the Tools:

All the tools that will be used are Python and it’s packages that are in anaconda such as:

* Pandas
* Seaborn
* Sklearn
* Matplotlib
* Numpy
* **MVB Goal:**

To present a notebook or PowerPoint presentation that shows the insights, model predictions, accuracy and results have been found during the creation of the model

Thank you for reading, see you at the notebook.