PHASE – 1: initial impression

1. Task: To determine which patients require more attention and resources from the healthcare system. To which patients should we divert resources.

2. Defining the problem: This is a classification problem. It aims to find out which individuals are likely to have 1 of 5 common cardio-vascular diseases, and which are not.

3. The Data: There are multiple anonymous medical records available, aggregating information about patients and their different medical outcome. Usually offered in the form of a table.

4. Data Size: My data is a 12x918 table with 11 features and 918 rows.

5. Age: Age of patient at checkup (years). [Numeric]
   Sex: Either Male or Female. [Categorical – M/F]
   Chest Pain Type: Type of chest pain reported by the patient. [Categorical – 4 types]
   Max HR: Maximum Heart-Rate reached during exercise (bpm). [Numeric]
   Exercise Angina: Chest pain during exercise. [Categorical – Yes/No]
   ST Slope: The slope of a particular area in the wave form of a heartbeat ECG during exercise. Correlates to different types of coronary artery problems. [Categorical – 3 types]

Unfortunately, datasets of this type seem to measure different features than the dataset I am using. Since this is individualized clinical data I also cannot add more features via feature engineering since I cannot apply measurement of particular individuals to others, that would render the interpretation of the data meaningless.

Phase 2 – Summary:

I managed to develop a model that receives a recall score of 91% and an f1 score of 92% for the category of 'Has a Heart Disease'.

It should be noted that a single unreliable data point caused the entire algorithm 'confusion' and made the scores drop by 2% each(For one patient the value of RestingEGC was 0 – a medical impossibility). This comes to stress out how important it is to include in future data sets only entire sets of features of each patient, and not fill the blank with random numbers such as 0.

Oddly enough, another feature which had many impossible values of 0 – Cholesterol levels – seems to not have create 'confusion' within the algorithms as removal of or tampering with the lines containing the values of 0 had negative impact on the performance of the algorithm – possibly because of the other features present and their correlation to the cholesterol factor.
Interestingly, training the algorithm on the full dataset yet testing in only on the rows with the cholesterol value other than 0 achieved significantly better results.
As such, it is hard to interpret the importance of this feature other than to imagine how better the diagnosis might be if that parameter was examined with every patient rather then filled with a blank 0 in cases where the test for cholesterol levels was not performed.

The efforts to reduce dimensions emerged with nothing, as it seems that no feature could be dropped without a severe reduction in precision for the 'Has a Heart Disease' category, and in severe reduction to all scores regarding the 'No Heart Disease' category, even though better recall score were achieved for the former.

This brings me to conclude that heart disease cannot be derived from any one factor present in the current research, only a holistic approach and the consideration of the combination of multitude of factors can indicate the presence of a heart disease, with no one factor seeming to be more indicative than the other.

It appears to be the case that in order to achieve better diagnosis, health professionals should strive to gather and consider as many physiological metrics as possible in order to determine an accurate diagnosis.

This brings to mind two possible ways of moving forward with the research:

1) To create a greater dataset of patients with complaints of chest pain with even more biological indicators, and if possible, more patients, to perhaps provide better training for the algorithm and arrive at better results which could in turn be used as a guiding tool for diagnosis in the future.

2) Use the knowledge gathered so far and try to search for bio-indicators which encapsulate the correlation between the features used so far. That is to say, find other physiological metrics that correlate to the features studied in this research – in order to perhaps further guide the field of research to find more accurate and specific bio-indicators for heart diseases, which would require less tests and perhaps include parameters which are more causative of the disease rather than correlative.