Gathering data.

Data requirements.

In order to accomplish our main goal of creating a working model which predicts whether a patient has heart disease or not based on a number of medical features, we will need a dataset of patients with their information regarding these features and whether or not they have heart disease. Preferably this dataset will be in CSV format for easier use. For this, we have decided to use a dataset from Kaggle which is in accordance with all of our requirements.

Verify data availability.

Kaggle datasets are publically available and easily downloadable, so there are no problems with data availability.

Define selection criteria.

The specific data source that we will use is the heart.csv file found within the Kaggle dataset. Within the file, every row and column is relevant to our project.

Describing data.

Within the heart.csv file there are 918 rows, each row containing a unique patient’s data. There are 12 columns, each column containing the value for a medical feature regarding the patient on that row. These features are:

* Age: the patient’s age.
* Sex: the patient’s sex.
* ChestPainType: the type of chest pain that the patient is experiencing. ATA (Atypical angina), NAP (Non-anginal pain), ASY ( Asymtomatic - Silent (asymptomatic) myocardial ischemia (SMI)) or TA (Typical angina).
* RestingBP: the patient’s resting blood pressure in mm Hg.
* Cholesterol: the patient’s total seerum cholesterol in mg/dL.
* FastingBS: the patient’s fasting blood sugar level, 1 if over 120 mg/dl (considered high), otherwise 0.
* RestingECG: the patient’s resting electrocardiogram results, Normal (normal), ST(having ST-T wave abnormality) or LVH (showing probable or definite left ventricular hypertrophy by Estes' criteria).
* MaxHR: the maximum heart rate achieved for the patient.
* ExerciseAngina: whether the patient has exercise-induced angina, Y(yes) or N(no).
* Oldpeak: a line in EKG doesn't come back to zero state, where it should come. ST [Numeric value measured in depression] ST depression induced by exercise relative to rest.
* ST\_Slope: the slope of the peak exercise ST segment (the shape of the line in a certain place in EKG). Up (upsloping), Flat (flat) or Down (downsloping).
* HeartDisease: whether the patient has heart disease or not, 1 for yes, 0 for no.

These features are enough for us to create our prediction model.

Exploring data.

The most obvious data quality problem is the fact that in this dataset there aren’t a lot of patients. 918 is a pretty low number of instances to create a good model, and considering that we also want to be able to find the features that are the most correlated to heart disease separately among male and female patients and within different age groups, then the number of instances in each of these groups can get really small. Also, the number of patients with heart disease is a lot smaller than the number of patients without heart disease. A model created with such skewed data would not be accurate, which is why we have to prepare the data by balancing it.

Verifying data quality.

The data we need exists, we can have it and there aren’t any fatal data quality issues. With proper data preparation, all the quality issues can be fixed and creating a good model will be possible. This single dataset is good enough to accomplish our goals.