Data exploration is the first step of performing any meaningful data analysis.

The data analysts can use both visual exploration tools and descriptive statistics to understand what are the characteristics of the data. Characteristics that are interested include:

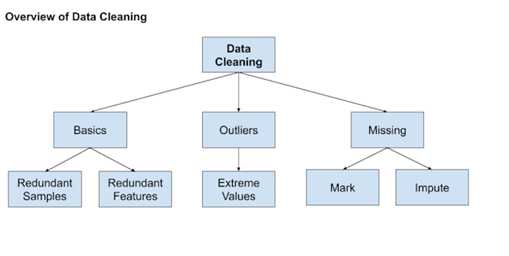
1. **the data size (both number of columns and number of rows);**

**Dataset**:

All four files of the dataset

Describe the input variables and output variables

* Cleveland Dataset (shape and head())
* Hungarian Dataset (shape and head())
* Switzerland Dataset (shape and head())
* VA dataset (shape and head())



1. **Data problems such as Basics, Outliers,missingness;**

* **Basics:** Redundant samples( zero rows)

Redundant Features (column with only one variable)

* **Missing values:**

1. **Mark:**

* Missing values( with ?/NA/ blank)---mark them as NAN
* Draw the histogram matrix
* Describe the data( once it is marked NAN then we can calculate)
* **Outliers:** plot boxplot for all input variables

1. **Imputation:**

KNN:(Cl, cl+sw, cl+hu, cl+sw+hu)--print the accuracy with std and k

Iterative: (Cl, cl+sw, cl+hu, cl+sw+hu)--print accuracy and std

**Decision: Final dataset for our project (cl+sw+hu with accuracy 0.857407 and std= 0.024038, k=3)**

**EDA of new dataset:**

* shape(), head(0, describe(),
* Outliers: extreme values

**(3) data distribution such as normalness, skewness, and kurtosis;**

* Histogram matrix (with color coded if possible)-- if skew then we can try transformation---in classification we need transformation on input only, am I right??
* class distribution to check the balance in the output variable

**(4) relationship among variables such as linear correlation, monotone correlation, and association between categorical variables.**

* Scatterplot matrix
* Correlation matrix to confirm the relation as seen in histogram

--------------------------------------------------------------------------------------------------------------------

**Feature selection: backward, forward, with some algorithms**

I think we did it in our lab session too..(also, homework assignment 1)

**Building a model:**

* Logistic regression
* Knn classifier
* SVM
* Generalized additive models
* Random forest
* Boosting

---

[Heart Disease Dataset](https://archive.ics.uci.edu/ml/datasets/Heart+Disease?spm=5176.100239.blogcont54260.8.TRNGoO)

\*\*Data Set Information:\*\*

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to

this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

To see Test Costs (donated by Peter Turney), please see the folder "Costs"

---

\*\*14 Attributes used:\*\*

age, sex, cp, trestbps,chol, fbs, restecg,thalach, exang, oldpeak, slope, ca, thal,num

\*\*Description of 13 Input variables\*\*

\*\*Age:\*\* age of the patient

\*\*Sex:\*\*

0: Female

1: Male

\*\*Chest Pain Type (cp):\*\*

0: Typical Angina

1: Atypical Angina

2: Non-Anginal Pain

3: Asymptomatic

\*\*Resting Blood Pressure:\*\* Person's resting blood pressure(trestbps).

\*\*Cholesterol (chol):\*\* Serum Cholesterol in mg/dl

\*\*Fasting Blood Sugar (fbs):\*\*

0:Less Than 120mg/ml

1: Greater Than 120mg/ml

\*\*Resting Electrocardiographic Measurement(restecg):\*\*

0: Normal

1: ST-T Wave Abnormality

2: Left Ventricular Hypertrophy

\*\*Max Heart Rate Achieved (thalch):\*\* Maximum Heart Rate Achieved

\*\*Exercise Induced Angina (exang):\*\*

1: Yes

0: No

\*\*ST Depression (oldpeak):\*\* ST depression induced by exercise relative to rest.

\*\*Slope:\*\* Slope of the peak exercise ST segment (slope):

0: Upsloping

1: Flat

2: Downsloping

\*\*Thalassemia (thal):\*\* A blood disorder called 'Thalassemia' :

0: Normal

1: Fixed Defect

2: Reversible Defect

\*\*Number of Major Vessels (ca):\*\* Number of major vessels colored by fluoroscopy.

\*\* 1 Output variables:\*\*

\*\*num:\*\* diagnosis of heart disease (angiographic disease status):

0- < 50% diameter narrowing

1- > 50% diameter narrowing

---

**How accurately can we predict the presence of heart disease using classification?**