Exploratory Data Analysis

Aayush Shrestha Anmol Jha

Kathmandu University
Department of Computational Mathematics

MATH 252 Progress Presentation
Supervisor: Mr. Kiran Shrestha

Outline

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a method of analyzing data using statistical summaries and graphical representations.

EDA is a critical process that aids in determining how to best manipulate data sources to obtain the required answers, making it easier for data scientists to discover patterns, detect anomalies, determine test hypotheses, and validate assumptions.

Dashboard

A dashboard is basically a GUI for Data Visualization. A dashboard is a good choice if you need to summarize and present a lot of information on a single window.

A dashboard helps Data Analysts and Data Scientists perform many data-related tasks and also provides a visual aid for other stakeholders to understand data and make accurate data-based decisions.

Goals for the Project

- Familiarity with data cleaning, feature extraction, and data conversion.
- Describing the data using statistical methods such as mean, median, standard deviation, correlation, and so on.
- Implementation of different plots in R.
- Implementation of a dashboard.
- Understanding and explaining the trends and outliers in the data based on the domain.

Initial View

SAMPLE DATA

Patient Id	PID0x81d5	*Test 1*	0	*Heart Rate (rates/min*	Normal
Patient Age	7	*Test 2*	0	*Respiratory Rate (breaths/min)*	Tachypnea
Genes in mother's side	Yes	*Test 3*	0	*History of anomalies in previous pregnancies*	Yes
Inherited from father	Yes	*Test 4*	1	*No. of previous abortion*	2
Maternal gene	Yes	*Test 5*	0	*Birth defects*	Singular
Paternal gene	Yes	*Parental consent*	Yes	*White Blood cell count (thousand per microliter)*	7.785072984
Blood cell count (mcL)	4.743537401	*Follow-up*	High	*Blood test result*	slightly abnormal
Patient First Name	Irene	*Gender*	Female	*Assisted conception IVF/ART*	Yes
Family Name	Trainer	*Birth asphyxia*	No record	*Symptom 1*	1
Father's name	Isaul	*Autopsy shows birth defect (if applicable)*	No	*Symptom 2*	1
Mother's age	31	*Place of birth*	Institute	*Symptom 3*	1
Father's age	61	*Folic acid details (peri-conceptional)*	No	*Symptom 4*	0
Institute Name	New England Medical Center	*H/O serious maternal illness*	No	*Symptom 5*	1
Location of Institute	818 HARRISON AV SOUTH END, MA 02118 (42.335925371008436, -71.07378404269969)	*H/O radiation exposure (x-ray)*	Not applicable	*Genetic Disorder*	Single-gene inheritance diseases
Status	Deceased	*H/O substance abuse*	Yes	*Disorder Subclass*	Cystic fibrosis

Figure: Sample raw data.

Initial Pre-processing

First of all, the data was changed from an xlsx file to a csv file, and the column names were modified to better suit programming. Also, all the rows with any NA values were removed.
After completing all that, the dataset now contains 45 columns and 6706 observations.

EDA Work

Exploring the columns and their types, the columns with data related to the patient but not the disorder were removed. Then, a statistical summary of the numeric columns was created, and the columns having no variance were also removed. We are left with 32 columns.
Furthermore, based on graphical methods such as stacked bar graphs, box and whisker plots, and domain knowledge, a few more columns were disregarded based on their distribution and implication.

Dashboard

We have a basic dashboard framework ready and we are just tweaking the data and the code to make it more robust for proper use.
In that dashboard, the column of numeric values shows its histogram and those with categorical value creates pie chart.

Remaining Work

- Develop a better dashboard with more and better visual aids and information either in R or in other frameworks.
- Further reduce the dimensionality of the parameters based on more robust statistical and domain knowledge.
- Finally, create a properly processed dataset to be used for future classification projects.

THANK YOU!

Please share if you have any queries.

We would like to show the code now.