CM 2nd Year 2nd Semester (MATH 252)

Exploratory Data Analysis (EDA) and Prediction on genetic data.

15th February 2023

OVERVIEW

Using the dataset provided here: kaggle-genetic, containing 43 features and 2 prediction columns. Apply descriptive statistical methods to find and present the trends that can be obtained from the data. The processed data will then be used for the dashboard.

Team: Currently individual

GOALS

1. Familiarity with the concept of Data cleaning, Feature extraction, Data conversion
2. Implementation of Different Plots using Packages provided in R.
3. Understanding and explaining the trends and outliers found in the data.
4. Implementation of a Dashboard (if possible also a classification prediction model) in R

SPECIFICATIONS

- **Tech:**
    - Language: R (tidyverse packages +Dashboard packages)
    - IDE (Jupyter notebook)

The Data deals with information related to whether a specific gene is present in the mother and or father gene. And also the medical data of the patient and result from at most 5 tests the last two columns contain the Genetic Disorder and genetic disorder subclass(for classification). Few of the data are strings, few will have to be converted from string to integer(yes, no to 1,0). The main objective is to visualize the predictor clas's relation with other features and possibly explain the reasoning behind the relationship between them.

Resources

- Exploratory Data Analysis in R for beginners (Part 1) | by Joe Tran | Towards Data Science
- Exploratory Data Analysis in R Programming - GeeksforGeeks
- How to Perform Exploratory Data Analysis in R (With Example) - Statology
- GGplot 101: The Ultimate CheatSheet | Kaggle
- Cheatsheet 70+ ggplot Charts 📊 | Kaggle
- Cool Dashboard in R with YouTube Tutorial | Kaggle
- How I create dashboard in R? | Data Science and Machine Learning | Kaggle
- Suicide data - Full interactive dashboard | Kaggle
- Interactive Dashboards in R | Kaggle