SRIHARSHA DAPARTI

UX46295

DATA 606 – CAPSTONE IN DATA SCIENCE

A VIEW ON HCV DATA

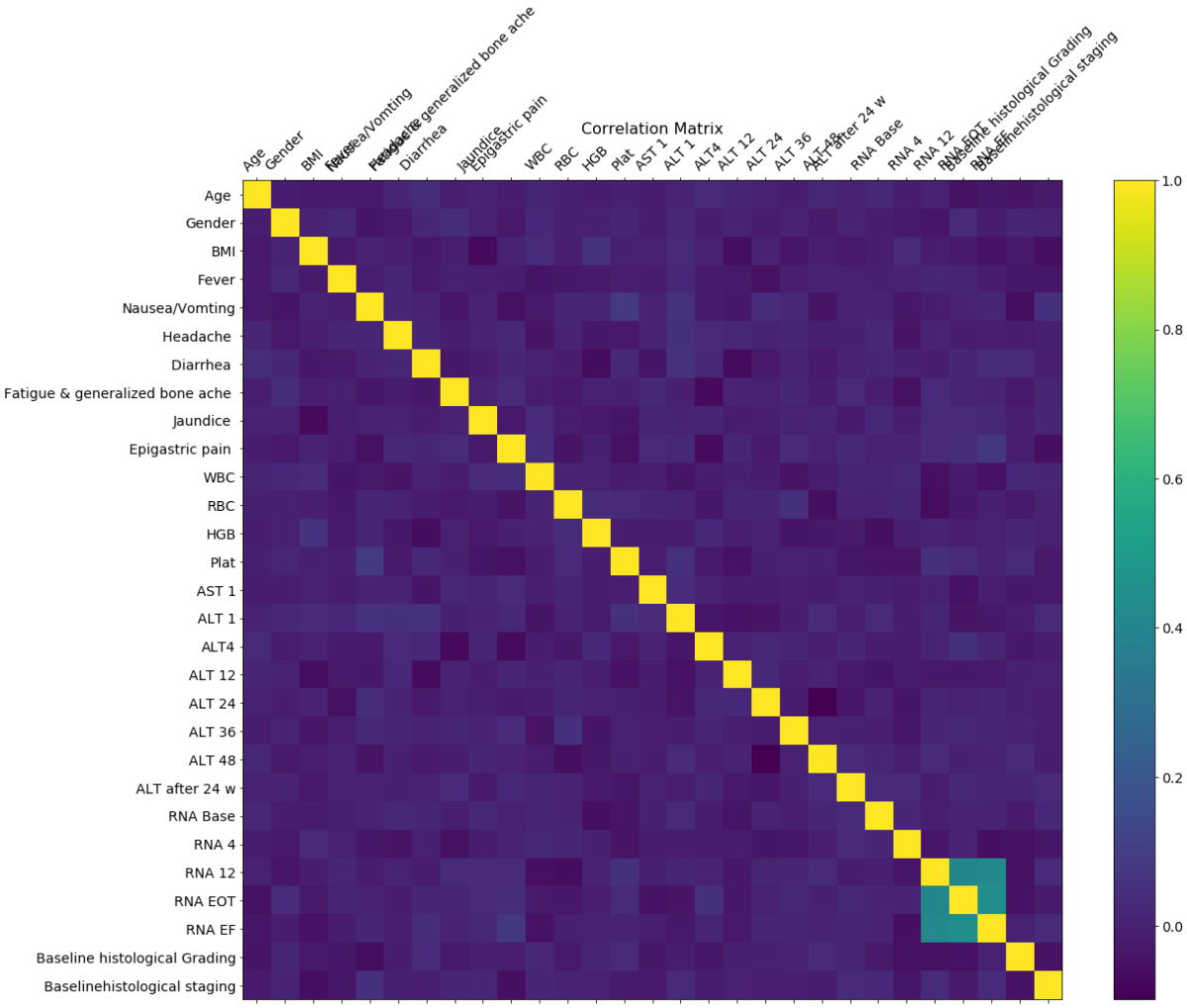
LITERATURE REVIEW

The research on how to tackle the presented problem and the quest for similar problems and the methodologies used to solve those problems was challenging. The primary paper[1] that used this data set was thoroughly examined along with the purpose of the research, approach and methodologies. The paper talks about coming up with 'Rules' which are combinations of two or more symptoms observed. The intention of this project is to take a more broadened approach. In this paper, the expected findings are the study of each individual symptom correlated with the test results.

Many other works[2] mentioning health data were also examined and a classification approach is decided to be the best way to tackle this question.

DATA CLEANING AND WORK DONE

Data Cleaning has been a very important and excruciatingly painful part of the project. Upon the review of the primary research publication that made use of this data set, it has been confirmed that the data set available at the data source has been mangled with. Working with such a dataset is very troublesome. At the end of about three weeks from the project decision, a repetitive and considerable amount of cleaning and viewing has been done on the data set for it to be fit for further analysis. Going further deep into the hurdles presented during this phase, discretization was a major chunk of the problem. Discretizing the attributes of the data set one by one, due to their varied nature has been an issue. With the help of a few supporting functions, a 'refined' discretized data set was produced. Following this cleaning process, the correlation between different attribute was visualized using a color coded Correlation Matrix plotted using 'matplotlib' package to get an idea of which direction this problem is supposed to be handled. This has shown that RNA levels at the end of 12 weeks and at the End of Treatment has shown to be playing an important role. The next step is to check which other attributes (symptoms) collide with this attribute to result in the said Baseline Histological Staging (whether the liver has Fewer Fibroids, Many Septa or is Cirrhotic).



REFERENCES

1. Nasr, M., El-Bahnasy, K., Hamdy, M., & Kamal, S. M. (2017). A novel model based on non invasive methods for prediction of liver fibrosis. 2017 13th International Computer Engineering Conference (ICENCO).
2. Kelman, C. W., Bass, A. J., & Holman, C. D. J. (2002). Research use of linked health data - a best practice protocol. Australian and New Zealand Journal of Public Health, 26(3), 251–255.
3. Patton, G. C., Coffey, C., Sawyer, S. M., Viner, R. M., Haller, D. M., Bose, K., … Mathers, C. D. (2009). Global patterns of mortality in young people: a systematic analysis of population health data. The Lancet, 374(9693), 881–892.