Prediction Model for Fibrosis Stages in HCV Patients

Ali Gamal
aly.elgaml98@eng-st.cu.edu.eg

Khaled Maher
khaled.bedda98@eng-st.cu.edu.eg

Nada Ashraf
Nada.Ibrahim98@eng-st.cu.edu.eg

1. Introduction

Hepatitis C is a serious and potentially life-threatening condition caused by the hepatitis C virus (HCV) that primarily affects the liver. During the initial infection people often have mild or no symptoms so it's possible that you may be infected for many years before you're diagnosed as the virus persists in the liver in about 75% to 85% of those initially infected. Possible symptoms include fever, dark urine, abdominal pain, and yellow tinged skin. Over many years however, it often leads to liver disease and occasionally cirrhosis. [1]

There is no vaccine against hepatitis C. An estimated 143 million people (2%) worldwide are infected with hepatitis C as of 2015. In 2013 about 11 million new cases occurred.[11] It occurs most commonly in Africa and Central and East Asia. About 167,000 deaths due to liver cancer and 326,000 deaths due to cirrhosis occurred in 2015 due to hepatitis C. [2]

Due to the large population of patients, our efforts will be concentrated on creating a prediction model using Machine Learning to correctly determine the stage of fibrosis of an HCV patient after an arbitrary period of treatment. This will mainly depend on various methods that will be outlined below in the Methodology section.

2. Motivation

Data science is an absorbing domain. Transforming raw data into meaningful insights is a powerful tool for biomedical engineers and their advancement of technologies. We have great enthusiasm for this project because it meets our research interests and will provide us a great opportunity to learn new skills and dive into the larger world of data science & analysis. Moreover, the data we will work on is collected from Egyptian patients and this relevancy makes us even more enthusiastic to fulfill our task.

3. Problem Statement

We will be building a machine learning model that is capable of predicting the stage of fibrosis that a HCV patient could catch after 48 weeks treatment.

4. Resources

Our work will be based on the dataset provided by the University of Ain Shams that includes data about Hepatitis C Virus (HCV) for Egyptian patients. Key attributes and possible values can be found in the link. EDA, Data Visualization, Pre-processing and modeling will be handled by the suitable R libraries.

5. Methodology

5.1. Exploratory Data Analysis

5.1.1 Data variation

Exploring patterns of variation, typical values and outliers is an important task. We can gain such knowledge by visualizing the variables' distributions. To examine the distribution of a categorical variable, we can use a bar chart. And for continuous variables, histograms and frequency polygons can be used. To overcome binning bias of histogram and display all data, we can use swarm plots.

5.1.2 Co-variance

It's important to study the behavior between variables to gain useful insights that can be useful for feature selection. To examine the covariance between categorical and continuous variables we can use a boxplot or violin plot. If both variables we are interested in are categorical we can use heatmap or scatterplot. If both are continuous, heatmaps can be used. Measuring the central tendency mean and median for numerical data and mode for categorical- and measuring spread of data. Examining relationships - plotting for numerical and two-way-cross-tabulations for categorical.

5.2. Data Pre-processing

5.2.1 Discretization

Discretization is the process of transforming continuous data into categorical data. The importance of discretization is that it helps handling outliers by placing these values into the lower or higher intervals together with the remaining inlying values of the distribution. Our discretization will be handled by the file attached to the dataset.

5.2.2 Feature scaling

Using a normalization technique (Z-score or min-max normalization) to avoid skew towards high magnitude features.

5.2.3 Feature Engineering

Categorical variables encoding and numerical variables engineering.

5.2.4 Feature selection

Through removing redundant features, checking for correlated features and training the model with feature selection and using PCA.

5.3. Modeling

We will be using both Decision Tree and Naive Bayes for our model and testing which one is more accurate. We may use LDA for dimensionality reduction. Should we have the time we may test further methods (not including binary methods).

5.4. Evaluation

Evaluation of our model will be done through comparing the actual outcome grading to the predicted class grading. In other words, our main measure of success will depend on the accuracy of our prediction model and it's ability to correct it's predictions with time.

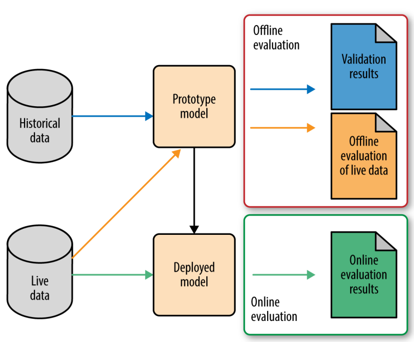

Figure 1. Machine learning model development and evaluation workflow [3]

5.5. Deployment

We plan on offering our predictions through an API to be designed later on once our algorithm is completed and working as planned.

6. Milestones and Contributions

Milestone	Date	Contributor
EDA	29 Oct - 2 Nov	Khaled Maher
Pre-Processing	3 - 8 Nov	Ali Gamal
Model Building	9 - 15 Nov	Nada Ashraf
Model Evaluation	16 - 18 Nov	Ali Gamal
Model Improvement	1 - 7 Dec	Khaled Maher
Deployment	8 - 10 Dec	Nada Ashraf
Documentation and Publicity	11 - 13 Dec	Ali Gamal

References

[1] C. G. Ray and K. J. Ryan, *Sherris medical microbiology: an introduction to infectious diseases*. McGraw-Hill, 2004.

[2] M. H. Forouzanfar, A. Afshin, L. T. Alexander, H. R. Anderson, Z. A. Bhutta, S. Biryukov, M. Brauer, R. Burnett, K. Cercy, F. J. Charlson *et al.*, "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet*, vol. 388, no. 10053, pp. 1659–1724, 2016.

[3] A. Zheng. (2015) Evaluating Machine Learning Models. [Online]. Available: https://www.oreilly.com/ideas/evaluating-machine-learning-models/page/2/orientation