Indian Institute of Information Technology, Allahabad
Software Engineering

Instructors: Dr. Sonali Agarwal

SOFTWARE REQUIREMENT SPECIFICATION

Change identification in data patterns Hepatitis data in India.

GROUP MEMBERS-

Medha Balani	**IIT2019021**
Vidushi Pathak	**IIT2019027**
Aarushi	**IIT2019032**
Jyotika Bhatti	**IIT2019036**

Table of Contents

1. Introduction

1.1 Purpose

Hepatitis is an inflammation of the liver. The condition can be self-limiting or can progress to fibrosis (scarring), cirrhosis or liver cancer. Hepatitis viruses are the most common cause of hepatitis in the world but other infections, toxic substances (e.g. alcohol, certain drugs), and autoimmune diseases can also cause hepatitis. There are 5 main hepatitis viruses, referred to as types A, B, C, D and E. These 5 types are of greatest concern because of the burden of illness and death they cause and the potential for outbreaks and epidemic spread. In particular, types B and C lead to chronic disease in hundreds of millions of people and, together, are the most common cause of liver cirrhosis and cancer (https://www.who.int/news-room/q-a-detail/hepatitis).

This software mainly analyses the previous trends in hepatitis disease in patients and according to those trends, it predicts the future critical conditions in patients, and how they can be prevented. This software can be used by hospitals and the department that deals with hepatitis patients. It may prove to be helpful to predict the occurrences of the disease in a particular age group, and analyze the severity of it in that age group or gender, or its dependence on various other factors. These data can be further changed and a new analysis can be obtained. The main objective is to create a predictive model which can help the physician to identify the prognosis and survivability of hepatitis patients based on the health features given.

1.2 Document Conventions

The workflow of the software development project is shown via a workflow diagram for the detection of the visualization of the dataset according to the different causes of the dataset values and the relation shown via the correlation matrix.

1.3 Intended Audience and Reading Suggestions

This document is mainly intended for the people in the medical field. They can analyze the data i.e the attributes that result in such a dataset and deal accordingly with hepatitis infection.

1.4 Product Scope

Provide all the necessary information and analysis about the trends of the disease hepatitis. Gives the information about attributes that bring in such trends, for example, age, gender, pre-medical history, etc. So, the scope of such a product is significant enough to help our medical sector. Further, the algorithm used in this model can come helpful while making software for other diseases.

The software will be **creating pop-ups** for change that will be identified in the data pattern. The **correlation matrix** will be used to identify the correlation coefficients and the way they will vary with each other, which will help in **summarizing the data** which is an input to the more advanced analysis and will make diagnosis easier. As the dataset will depict all the **previous records** of the Hepatitis Virus , this will make the creation of models more efficient and the computation of the metrics will also become easier. The mechanism used in the software will be such that .

1.5 References

- https://www.who.int/news-room/q-a-detail/hepatitis
- https://datahub.io/machine-learning/hepatitis
- https://www.kaggle.com/harinir/hepatitis

2. Overall Description

2.1 Product Perspective

This software is aimed to help people analyze the change in pattern of hepatitis in India. They can see the changes in the **pattern of hepatitis according to various attributes like age, gender, antivirals etc.**

2.2 Product Functions

This software shall perform the major functions on its own according to certain *algorithms designs* and shall provide the user with the required information. It will accordingly study the available data and come out with the results that the user may need in order to treat the patient.

2.3 User Characteristics

The user should be familiar with the operation of web applications.

2.4 Operating Environment

This software can be run on any operating system which should have python installed in it and set in the environment path variables. This is a web based application built in python so the required libraries such as **pandas, matplotlib, keras** etc and **.exe file** would be provided in the git repository of this web based app and can be directly accessed from there.

2.5 Design and Implementation Constraints

The data taken here being static in nature, users can't change the type of analysis embedded within the system of this application. Instead they can get an overall analysis report according to the previous trends. The application won't be functional if the system does not fulfil the operating environment of the software. Here we have made use of

1.) **Python(Tkinter), for creating a GUI,**
2.) **Python, as a backend scripting language**
3.) **MySQL as database, for storing raw data**
4.) **Matplotlib, for dividing data into chunks on various grounds**

Here we will use communication protocol as **HTTP requests**. We won't be taking any personal data of users in any form. So any kind of authentication isn't required. Thus no security consideration.

We are planning to use **Pandas library** of python to analyse the data, which is a Python package that provides **fast, flexible, and expressive data structures** designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. So the use of pandas will help in the analysis and manipulation of the data.

Additionally after the manipulation of the data , we will make a visualization using the matplotlib, which is a comprehensive library for creating static, animated, and interactive visualizations in Python. **Matplotlib** produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. *Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.*

2.6 Assumptions and Dependencies

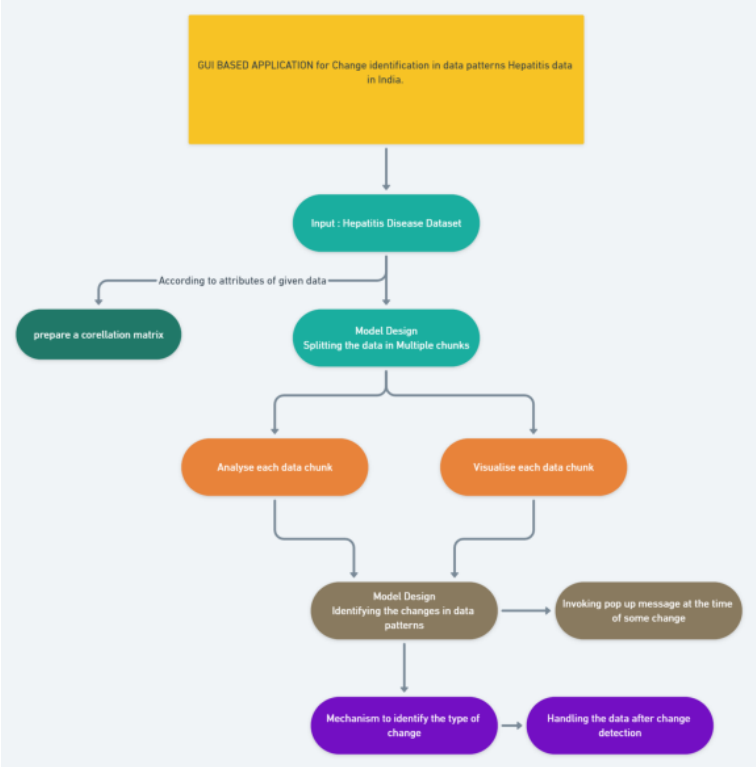
Here we have assumed the data sets taken to be of indian patients who are alive. Another important dependency is that these data are not confidential, or personalised. The information taken here is extracted from some existing project.

3. External Interface requirements

3.1 User Interfaces

The software will be a GUI based application which will help in the visualization of the data chunks that will be splitted and analysed by the python scripting from the given data set. Then accordingly it will depict the visualization of each cause of the ‘hepatitis virus’ which will also create a pop up message of the entire information as soon as a change will be depicted by clicking the graph the user wants a detail of .

The correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.



3.2 Hardware Interfaces

There are no specific hardware interface requirements. Any, which has the ability / specifications to store required libraries and tools works.

3.3 Software Interfaces

The software will be making use of scripting language python which will be using the pandas and matplotlib for the analysis , manipulation and plotting and the visualization of the data patterns and the dataset will be stored as .csv in the backend of the software.

3.4 Communication Interfaces

This software will use communication protocol as HTTP requests and will give a pop up message to the user at the time of changes in data patterns.

4. System Features

4.1 Static data visualization

The dataset will be splitted into chuks and will be visualised according to the patterns observed on various factors . Also, we will be making the corresponding correlation matrix of every cause that has a major effect on it. The correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. The frequency plot table with age and other factors will be plotted along with

Dataset		Total observations		Total Features	
Hepatitis Domain Dataset		155		20	
Continuous vars	**Min**	**Mean**	**Median**	**Max**	**SD**
AGE (155)	7	41.2	39	78	12.57
BILIRUBIN (149)	0.3	1.43	1.0	8.0	1.21
PROTIME (88)	0	61.85	61	100	22.86
ALBUMIN (139)	2	3.82	4	6.4	0.65
ALK PHOSPHATE (126)	26	105.32	85	295	51.51
SGOT (151)	14	85.89	58	648	89.65
Categorical vars		**Details**			
SEX		Male: 139, Female: 16			
STEROID		No: 76, Yes: 78, Missing: 1			
ANTIVIRALS		No: 24, Yes: 131			
FATIGUE		No: 100, Yes: 54, Missing: 1			
MALAISE		No: 61, Yes: 93, Missing: 1			
ANOREXIA		No: 32, Yes: 122, Missing: 1			
LIVER BIG		No: 25, Yes: 120, Missing: 10			
LIVER FIRM		No: 60, Yes: 84, Missing: 11			
SPLEEN PALPABLE		No: 30, Yes: 120, Missing: 5			
SPIDERS		No: 51, Yes: 99, Missing: 5			
ASCITES		No: 20, Yes: 130, Missing: 5			
VARICES		No: 18, Yes: 132, Missing: 5			
HISTOLOGY		No: 85, Yes: 70			
CLASS (Outcome)		**Die: 32 (26%), Live: 123 (74%)**			

As mentioned in the table above, the dataset consists of 19 features and 1 Class (outcome), which can be categorized into 5 categories as below:

Category	Feature
Patient profile	SEX, AGE
Sign and Symptom	FATIGUE, MALAISE, ANOREXIA
Physical Examination	LIVER BIG, LIVER FIRM, SPLEEN PALPABLE, SPIDERS, ASCITES, VARICES
Lab Test	BILIRUBIN, PROTIME, ALBUMIN, ALK PHOSPHATE, SGOT, HISTOLOGY
Medication Management	STEROID, ANTIVIRALS

It is noticed that the **class is imbalanced** which consist of **26% of the patient die and 74% of the patient alive**. Imbalanced data in a classification problem possess a significant challenge in the quality of results obtained through the predictive models.

4.2 Dynamic Data Visualization

From the data to be input from the user about his details that may directly or indirectly affect the cause of the disease, we have to predict the probability of the diagnosis of the disease and how and at what rate it will affect also upto what extent . Along with the mortality rate.along with the age group classification also .

5. Other Nonfunctional Requirements

5.1 Performance Requirements

5.2 Safety Requirements

As, we are dealing with a very sensitive topic as it is related to the medical field. While predicting anything and providing info to users, we would want it to be correct and authentic. So any symptoms related to the diseases, links to blogs or anything would be authentic and also we are predicting the possibility of this disease in a person so we would try to use the most accurate and reliable dataset available.

5.3 Security Requirements

In this project there are not any specific security/privacy issues as we are not taking any personalised user data nor are we dealing with any kind of user authentication. Any user can put some details to fetch the prediction according to the data set analysis.

5.4 Software Quality Attributes

The definition of a quality software is basically its response time, so the time for fetching the raw data or displaying models or visualizations would be minimized. And also predicting probability etc. would be fast and as accurate as possible.

5.5 Business Rules

This project gives insight into the medical profession that's valuable in the world of data science . The medical professionals with data science knowledge will be helpful in knowing what features to add and what all major features and factors that determine the visualization.