University of Macau

Faculty of Science and Technology

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Visualizing Decision Rules and Relations for Medical Data Analysis: Modeling

by

LIANG LIHENG, Student No: DB626233

Final Project Report submitted in partial fulfillment
of the requirements of the Degree of
Bachelor of Science in Computer Science

Project Supervisor

Prof. Simon FONG & Prof. Shirley SIU

02 June 2020

DECLARATION

I sincerely declare that:

1. I and my teammates are the sole authors of this report,
2. All the information contained in this report is certain and correct to the best of my knowledge,
3. I declare that the thesis here submitted is original except for the source materials explicitly acknowledged and that this thesis or parts of this thesis have not been previously submitted for the same degree or for a different degree, and
4. I also acknowledge that I am aware of the Rules on Handling Student Academic Dishonesty and the Regulations of the Student Discipline of the University of Macau.

Signature : _____

Name : LIANG LIHENG

Student No. : DB626233

Date : 02 June 2020

ACKNOWLEDGEMENTS

The author would like to express his utmost gratitude to UM for providing the opportunity to carry out a project as a partial fulfilment of the requirement for the degree of Bachelor of Science.

Throughout this project, the author was very fortunate to receive the guidance and encouragement from his supervisors Prof. Simon FONG & Prof. Shirley SIU.

ABSTRACT

Visual techniques are very useful in data exploration because of the phenomenal abilities of the human visual system to detect structures and relations in images.

This is sometimes called visual data mining or visual mining that makes learning from visually presented information faster and is quite useful in exploration stage when the exact prediction target may be not very well known or needs to be confirmed preliminarily via studying the abstract data graphically.

In this project, we will divide into two part: machine learning & visualizing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. Introduction

1.1 Overview

1.1.1 Background

Medical diagnosis is a main part of in medicine. It is the first and important procedure to save the patient's life and improve health. As many possible reasons are involved, it is not easy for medical diagnosis and identify the cause of the illness. Traditionally, a doctor can only give a diagnosis result based on his/her knowledge and experience. Mentioned in [1], many researchers combine computer science and medical science in recent years. By taking advantage of computer science, medical research can be efficiently. Although computers cannot completely replace doctors for medical diagnosis, computers provide doctors with useful tools to help doctors make diagnosis efficiently and accurately.

'The invention of wearable sensors that can easily collect data on a person's day to day vitals makes machine learning systems incredibly useful in the health care industry. Decision tree style algorithms tend to be especially helpful. Patients can be broken down by group to track how certain variables affect the efficacy of treatment plans, or the likelihood of developing certain diseases. It can also help to track abnormalities in a given patient's health and use what it's learned from previous cases to predict what the cause of the abnormality might be.'[2]

'There is also great value to be gained from machine learning at the clinical trial stage of medical procedures. The ability to quickly create accurate models based off of real-world data can reduce the number of physical trials that need to be run. This can speed the process of getting drugs from development to implementation and reduce the risk factor to patients.'[2]

1.1.2 Research problem

Using computer technology in medical data analysis accelerate medical progress. Nowadays, artificial intelligence and mechanical learning are already mature. By taking advantage machine learning, computer can generally provide a reliable and accuracy result. The traditional machine learning model is a black box learning process, and this technology sometime lacks interpretability. In medicine, interpretability is important for doctors to analyse and interpret data. It enables doctors to determine causal relationship between diseases, living habit, and other features.

Finding causal relationship is not enough, an intuitive representation of causal relationship is also required. A graphical representation is preferred because humans have the strong power to analyse graph. Data and graph are two different representation of information, converting data to graph is a challenge. No universal visualization tool is existed. Depending on the application, special visualization tool needs to be developed.

1.1.3 Importance

Visualizing data is important in data analysis. Humans have the strong power to analyse structures and relationships by studying graphical data. It may facilitate data

mining that learning from visually presented information may faster. It is useful in the stage of exploration if exploring goal is not well known. It may also help people to confirm preliminarily knowledge quickly by studying the graphical data.

1.1.4Motivation

This final year project is focused on analysing medical data. Our motivation is providing useful information for doctor or medical research by taking the advantage of computer power. Data information is not intuitive enough, providing a better result by using visualizing technology is also our motivation. By the advantage of the phenomenal abilities of the human visual system, visualizing result can improve understandability, usability and efficiency.

1.2 Objectives

1.2.1Goal

This approach is quite opposite to formal methods of model building and testing, but it is ideal for searching through data to find unexpected or unusual relationships. Therefore, the objective of this FYP is focused on a vertical solution for analysing medical data, which joins advantages of several data mining techniques in one system, which consist of following parts: Data recognition-this subsystem transforms raw data to a form suited for further data processing. Additionally, noise and redundant data are removed based on a statistical analysis. Feature subset selection which is responsible for selecting an optimal set of attributes for a clean generation of decision rules Rule induction subsystem which uses both classical machine learning algorithms as well as new ones to be proposed. Visualization of the collected knowledge in a form easily understandable by humans. In this FYP we extend some initial research on the data visualization on both cancer and liver malfunction testing data, which are real from a local and international hospital.

In addition, many of the computer technologies we learned in our bachelor's degree programs are useful for analysing medical data, including data processing, machine learning and visualization. Therefore, we would like to apple those technologies in this final year project. Data processing can deal with raw data and transform it into research-worthy data. Machine learning enables us to learn data and generate useful prediction. Visualization provides a concise, understandable and user-friendly representation to demonstrate the final result. We want to design a system that combines the advantages of different computer technologies which we have already learned to produce useful medical results.

1.2.2How to achieve

In this project, we divide it into two part: machine learning & visualizing. To find out and understand relationship between behaviours and the disease, we try to use algorithms like Bayesian network, decision tree etc. Besides, to make the result more ‘user friendly’ we are going to make some improvement through existed visualizing tool.

For the machine learning part, the main workflow is analysing the data to generate a useful result for the visualization part. A set of medical data is given. The data was collected by other medical researchers and published on the Internet. The data includes patients living habits, diseases and other factors. Then, data processing

methods can be used to filter out the invalid data. By using machine learning method, the relationships between different attributes can be found.

For the visualization part, the main task is to develop a visualization software to show the result of the machine learning. The result of the machine learning part is given by another sub team. Based on the result, a concise, understandable and user friendly graphical representation should be provided. Since this software is designed for doctors, useful functions for doctors should be considered.

CHAPTER 2. Literature Survey/Related Work

2.1 Machine learning for medical diagnosis

Medical diagnosis is an important part in medical situation. A precise diagnosis can save the human life and improve their health. However, a wrong diagnosis may lead to serious consequences.

As described in [1], many medical institutions use computer information tools for diagnosis now. Due to technical limitations, they are not a substitute for a doctor to make a direct medical diagnosis now. However, they support doctors by selecting or generate related data for the medical diagnosis. They make medical diagnosis procedure to be more efficient and accurate.

Mentioned in [1], machine learning is an effective technology for medical diagnostic systems. Bayesian networks are one of these methods. They are a graphical model that includes a set of variables and their probabilistic independences. They are always used to handle uncertain problems. They can be used for medical decision making process. In medical diagnosis procedure, they can predict the likelihood of a patient's illness based on detected symptoms.

Also mentioned in [3], Bayesian networks are graphical presentation of relationships between different variables. In Bayesian networks, the nodes represent variables. The arcs between different variables represent causal, influential, or correlated relationships. The structure of the Bayesian networks is very suitable for medical decision making.

In [1], the author introduced the four case studies based on recorded medical evidence and statistics from Lugoj Municipal Hospital. Author proved that Bayesian networks are an efficient tool for the doctors to predict and treat disease.

Author of [3] said that most of the real-world problems involve uncertain variables and data, especially for medical related problems. These problems are perfectly represented by Bayesian network. Thus, Bayesian networks caught people's attention. Application of Bayesian networks are now popular in solving medicine problem. They are been known as a powerful tool for risk analysis and decision support in real-world.

2.2 What is Bayesian network

Bayesian networks are a kind of graphical model. They can be used for representing the dependence relationship between a given set of random variables as a directed acyclic graph. According to the definition in the book “Bayesian Networks with Examples in R” [5], Bayesian networks graph can be defined as

$G = (V,A)$

They include two parts, a set of nodes V and a set of arcs A. An arc is either an ordered pair or a non-ordered pair. Generally, an arc is called directed arc if it is an ordered pair. Otherwise, it is called undirected arc. Each arc can be defined as

a = (u,v)

It represents that the arc is outgoing for u and it is incoming for v (u→v) if it is a directed arc. Otherwise, it is represented with a simple line (u-v). Depend on characterization of arcs, a graph can be defined as directed graphs if all arcs are directed, undirected graphs if all arcs are undirected, or partially directed graph if it includes both directed and undirected arcs (see Figure 1 and Figure 2) (Figures are from [4]).

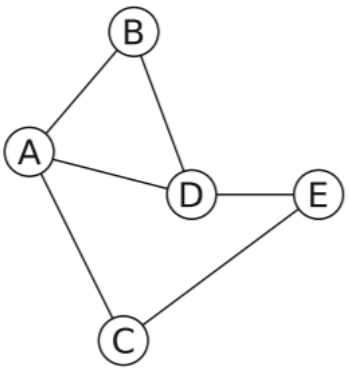

Figure 2-1: Example of undirected graph

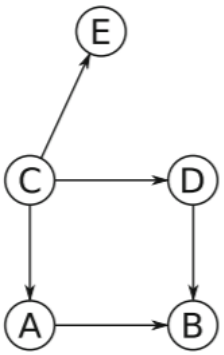

Figure 2-2: Example of directed graph

Different to other type of graph, Bayesian networks focuses on probability. They assign a probability to each measurable set of events. For discrete case, it is represented as conditional probability tables. This is the most common in the real-world application. For continuous case, it is represented as linear models [4].

2.3 Advantages and difficulty of Bayesian network

Bayesian networks are a powerful method to model uncertain variables and data. Thus, they become a popular tool for the analysis of uncertain problems. They provide a lot of many advantages to solve scientific problems. However, nothing is perfect, Bayesian networks also have some disadvantages.

They are the key points of the Bayesian networks. Author of [5] said that they provide a probabilistic representation of the interaction between different nodes. They make use of probability to measure uncertain data. If the uncertainty is higher, probability distribution is wider. As the amount of information increases, probability distribution

become narrower, it means uncertainty become lower. They allow users to better estimate risks and uncertainties. They have strong data analysis capabilities in areas full of uncertainty. They have the ability to convert uncertain problems to be possible. Therefore, they are often used in medicine and medical research because this type of research is always full of uncertainty.

According to [7], the second advantage of Bayesian networks is that they create a causal relationship between variables. Traditionally, people use black-box machine learning methods to do the research. It generally accepts a set of input variables, then it returns a result. There is hidden layer in the black-box machine learning methods. It is difficult to know what the hidden layer is doing. Thus, causal relationship between variables cannot be provided by black-box methods.

As mentioned in [7], Bayesian networks are white-box methods. They provide a direct presentation of the uncertain interactions between causes and effect. White-box methods allow incorporate domain knowledge in the model. Having representation of the uncertain relationships are useful for solving real world problems, such as including diagnosis, forecasting and information retrieval. For example, this help doctors and research to explain the disease.

In many other areas of research, dataset usually includes continuous data. As shown in [6], Bayesian networks can handle continuous data, but they have less support on continuous variables than other machine learning method. Generally, people solve this problem by converting continuous variables to discrete variables. This means that they discretize variables during the pre-processing data phase.

This is a trade-off. After discretization, the data can only retain the rough features of the original distribution. If too many characteristics of the original data are missing, it is impossible to show the exact relationship between variables. Discretization is an important task that can affect results. Discretization should carefully consider intervals and breakpoints, try different combinations to find the best combination. This is a time-consuming task.

2.4 Available tool for Bayesian network learning

2.4.1R and its libraries

R [11] is one of the programming languages and environments for computing and graphics. Similar to Python, another well-known programming language, it also has great library support for machine learning. It can be used to develop machine learning and analyse data, and to solve problems in the real world. Under the terms of the Free Software Foundation's GNU General Public License, it is free that everyone has the rights to freely use, study and modify it. In other words, it can be freely used for academic research of students.

According to [11], compared with other programming languages, R provides more extensive support for graphical techniques. Good at providing high-quality graphical solutions. The key to Bayesian networks is to provide a graphical representation of the data analysis. R is a suitable choice for developing Bayesian network learning.

According to [4], many people have developed useful libraries for Bayesian network implementations because of the advantage of the R language. Libraries bnlearn, deal, pcalg and pcalg are popular libraries dealing with Bayesian networks. They

respectively provide different functions. The main different of them have been showed in the following table (see Table 1).

	bnlearn	**deal**	**pcalg**	**pcalg**
Processing discrete data	Yes	Yes	Yes	Yes
Processing continuous data	Yes	No	Yes	Yes
Constraint-based learning	Yes	No	No	Yes
Score-based learning	Yes	Yes	Yes	No
Parameter estimation	Yes	Yes	Yes	Yes
Prediction	Yes	Yes	No	No

Table 2-1: Comparison of built-in libraries

Most of them provide basic structure learning algorithms and parameter learning approaches for both discrete and continuous data. They are the basis for implementing Bayesian networks.

Based on them, implementations can be faster and more efficient. These features can be used to build our own systems instead of writing everything from scratch. According to software engineering principles, save the cost of implementing basic functions as soon as possible, and focus on the main goal, which is to implement Bayesian network learning for medical data analysis. More time resources can be used to optimize the results, more suitable solutions for medical data analysis can be provided.

2.4.2Bnlearn

bnlearn is an R package for learning the graphical structure of Bayesian networks, estimate their parameters and perform some useful inference. It was first released in 2007, it has been under continuous development for more than 10 years (and still going strong). [8] In this period, we use ‘Hill Climbing’, a score-based structure learning algorithm from bnlearn to generate the plot.

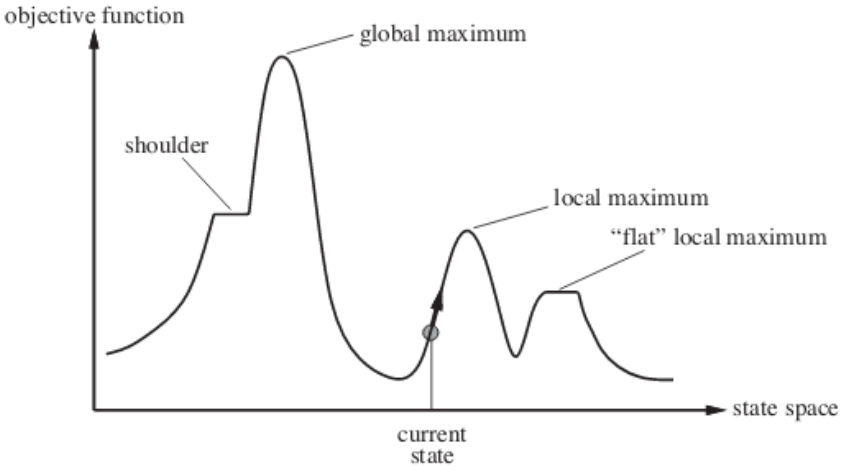

Figure 2-3: Example of Hill Climbing graph

Algorithm 1: Structural learning of BNs by using a hill climbing (HC) algorithm [9]

Input: D: A dataset defined over variables V = {X1,..., Xn}

Input: G0: A DAG defined over V used as the starting point for the search

Output: A DAG G being the graphical part of network B

1 G←G0;
2 fG←f(G:D)//*
3 [r]f: decomposable scoring metric improvement←true;
4 **while** improvement **do**
5 improvement ← false;
//*
6 [h]neighbors generated by addition
7 For each node Xi and each node X j ∈/ PaG(Xi) such that X j → Xi does not introduce a directed cycle in G, compute the difference diff = f(G+{Xj → Xi}: D)− fG. Store the change which maximizes diff in ⟨$change_a$, $diff_a$⟩;
//*
8 [h]neighbors generated by deletion
9 For each node Xi and each node X j ∈ PaG(Xi), compute the difference diff = f(G−{Xj → Xi}: D)− fG. Store the change which maximizes diff in ⟨$change_d$, $diff_d$⟩;
//*
10 [h]neighbors generated by reversal
11 For each node Xi and each node X j ∈ PaG(Xi) such that reversing X j → Xi does not introduce a directed cycle in G, compute the difference diff = d1 + d2 where d1 corresponds to f(G−{Xj → Xi}: D)− fG andd2 corresponds to f(G+{Xj → Xi}: D)− fG.Store the change which maximizes diff in ⟨$change_r$, $diff_r$ ⟩;
//*
12 [h]Checking if improvement
13 Let d* = max_k= $_{a,d,r}$$diff_k$ and move* its corresponding change;
14 if d* > 0 then
15 improvement ← true;
16 G ← apply move* over G;
17 fG ← fG + d*;

18 end
19 end

20 return G;

Hill Climbing is a heuristic search to find out the optimal result in the function, the main idea is to compare the point next to the current point and determine the continuous direction. In our project we use it to learn a network structure to attain relationship of different attribute from the dataset.

2.4.3Bayesialab

BayesiaLab is a powerful desktop application with a sophisticated graphical user interface, which provides scientists a comprehensive “laboratory” environment for machine learning, knowledge modelling, diagnosis, analysis, simulation, and optimization. With BayesiaLab, Bayesian networks have become practical for gaining deep insights into problem domains. BayesiaLab leverages the inherently graphical structure of Bayesian networks for exploring and explaining complex problems. [10]

BayesiaLab is the result of nearly twenty years of research and software development by Dr. Lionel Jouffe and Dr. Paul Munteanu. In 2001, their research efforts led to the formation of Bayesia S.A.S., headquartered in Laval in northwestern France. Today, the company is the world’s leading supplier of Bayesian network software, serving hundreds major corporations and research organizations around the world. [10]

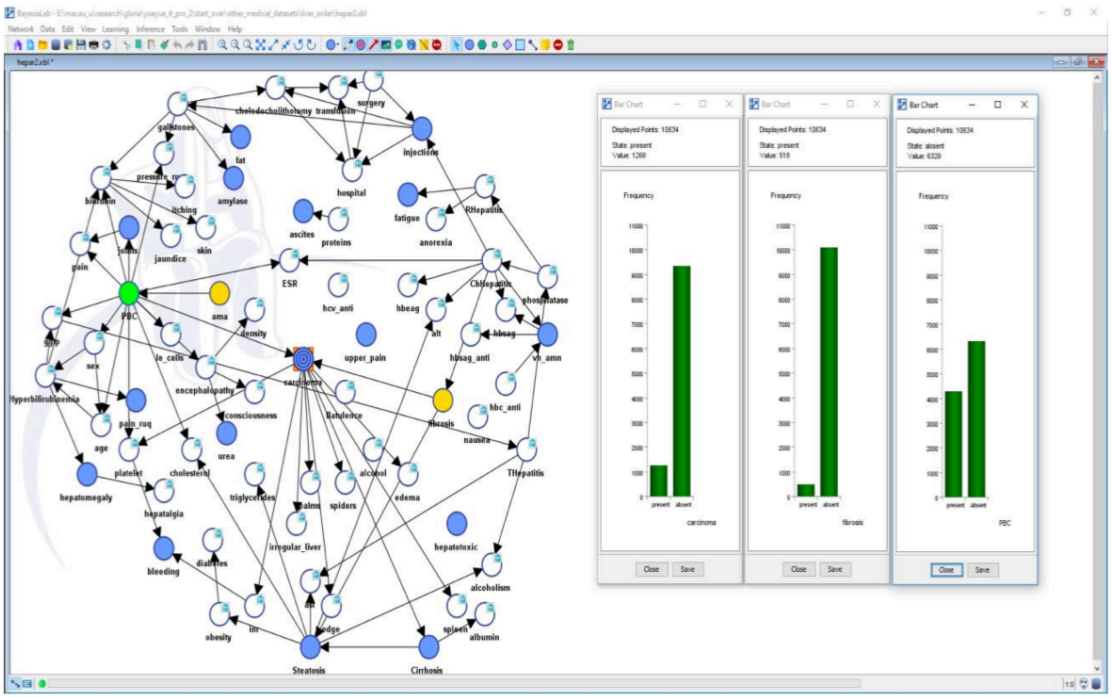

Figure 2-4: Example of Bayesialab result

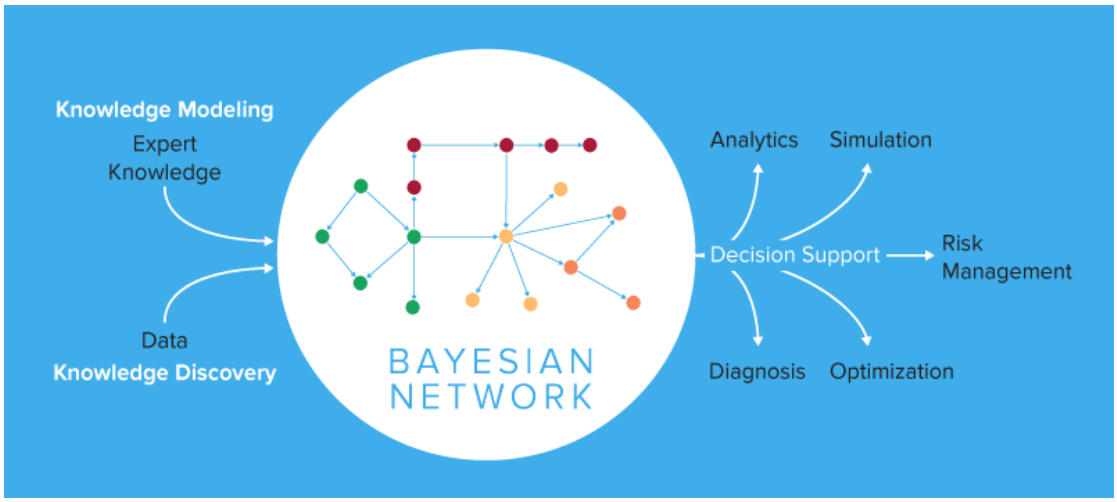

Figure 2-5: Example of Bayesian Network workflow

BayesiaLab is designed around a prototypical workflow with a Bayesian network model at the center. BayesiaLab supports the research process from model generation to analysis, simulation, and optimization. The entire process is fully contained in a uniform "lab" environment, which provides scientists with flexibility in moving back and forth between different elements of the research task.

However, we didn't get the license of BayesiaLab, we give up of this software.

2.5 Decision Tree

Decision tree is a method of machine learning. A decision tree is a tree structure in which each internal node represents a judgment on an attribute, each branch represents the output of a judgment result, and finally each leaf node represents a classification result.

Decision tree is a very common classification method that requires supervised learning, supervised learning is to give a bunch of samples, each sample has a set of attributes and a classification result. Then, by learning these samples, a decision tree is obtained, which can correctly classify the new data.

Information Gain:

$$IG(S,A) = H(S) - H(S,A)$$

Information gain is also called as Kullback-Leibler divergence denoted by IG(S,A) for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables.

Alternatively,

$$IG(S,A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

where IG(S, A) is the information gain by applying feature A. H(S) is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature A, where P(x) is the probability of event x.[12]

2.5.1'J48' Algorithm

We mainly use the 'J48' decision tree algorithm in Weka. J48 is an algorithm based on the 'top-to-down' and 'divide-and-conquer' recursively strategy. It will firstly select an attribute to be placed at the root node, then generate a branch for each possible attribute value, and divide the instance into multiple subsets, each of which corresponds to a root node branch. Then, to repeat this process recursively on each branch. When all instances have the same classification, the algorithm will come to an end.[13]

$$Info(D) = Entropy(D) = -\sum_{j} p(j \mid D) \log p(j \mid D)$$

$$Info_A(D) = \sum_{i=1}^{v} \frac{n_i}{n} Info(D_i)$$

$$Gain(A) = Info(D) - Info_A(D)$$

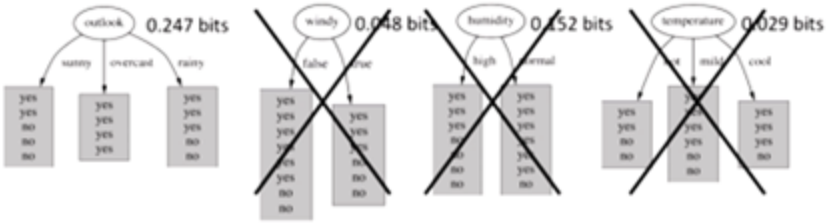

Figure 2-6: Sample 1 of J48

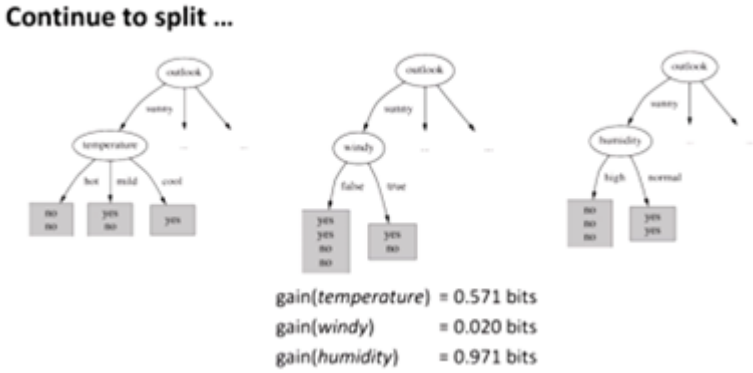

Figure 2-7: Sample 2 of J48

After calculating the gain, we compare it and select the smallest one as a root.

2.6 FURIA algorithm

FURIA (Fuzzy Unordered Rule Induction Algorithm) is a new fuzzy rule-based classification method, which is a modification and extension of the state-of-the-art rule learner RIPPER (Cohen, 1995).

FURIA learns fuzzy rules instead of conventional rules and unordered rule sets instead of rule lists. Moreover, to deal with uncovered examples, it makes use of an efficient rule stretching method. Fuzzy rules are more general than conventional rules and have a number of advantages. [14]

Suppose that fuzzy rules $r_1^{(j)}$... $r_k^{(j)}$have been learned for class λ_j. For a new query

instance x, the support of this class is defined by

$$s_j(x) \stackrel{\text{df}}{=} \sum_{i=1\ldots k} \mu_{r_i^{(j)}}(x) \cdot CF\left(r_i^{(j)}\right)$$

where $CF(r_i^{(j)})$ is the certainty factor of the rule $r_i^{(j)}$. It is defined as follows:

$$CF\left(r_i^{(j)}\right) = \frac{2\frac{|D_T^{(j)}|}{|D_T|} + \sum_{x \in D_T^{(j)}} \mu_{r_i^{(j)}}(x)}{2 + \sum_{x \in D_T} \mu_{r_i^{(j)}}(x)}$$

CHAPTER 3. Project Execution Schedule

The following tables showed that our schedule:

Table 3-1: Schedule in first semester

Task	Month			
	September 2019	October 2019	November 2019	December 2019
Grouping and selecting topic				
Learn the basics of Bayesian network				
Literature survey				
Finding available software				

The following table showed that our schedule in second semester:

Table 1-2: Schedule in second semester

Task	Month			
	January 2020	February 2020	March 2020	April 2020
Documenting interim report				
Design and implantation				
Testing and evaluation				
Documenting final report and preparing presentation				

CHAPTER 4. Functional Specification

4.1 Description of data

There are two datasets ‘HEPARTWO10k’ and ‘tox21’ for our project. In this case, ‘HEPARTWO10k’ is a dataset with 70 attributes and only the one ‘carcinoma’ is the target label representing if the liver cancer present or not.

Besides, ‘tox21’ is another dataset with 13 attributes, there are 12 target values to predict toxic or not, the other one is the structure of the molecule.

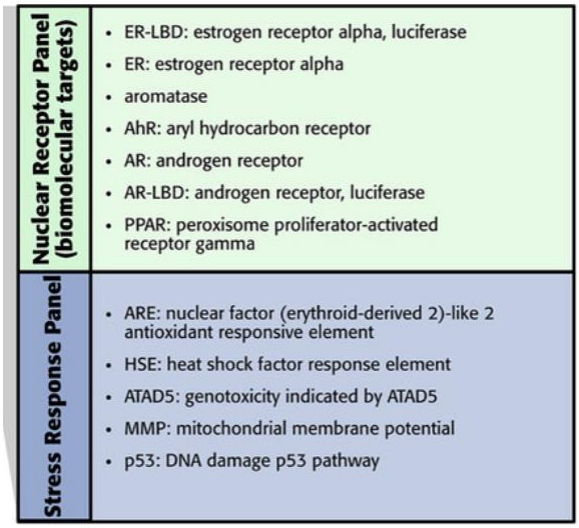

Figure 4-1: 12 Target values of dataset ‘tox21’

Our supervisor provide us a method to convert the structure into numeric attributes that help do the machine learning.

```
import deepchem as dc
import numpy as np
import pandas as pd
from rdkit import Chem

# load data as pandas DataFrame
dataset_file="tox21.csv"
df = pd.read_csv(dataset_file)

# get all molecule smiles from the smiles column
mols = [Chem.MolFromSmiles(smile) for smile in df['smiles']]

for i, s in df['smiles'].items():
    print(i, s)

# featurize all molecules
feat = dc.feat.RDKitDescriptors()
matrix = feat.featurize(mols)
featname = dc.feat.RDKitDescriptors.allowedDescriptors

# create a DataFrame
newdata=pd.DataFrame(data=matrix,columns=featname)

# append the mol_id
newdata.insert(0,'mol_id',df['mol_id'], True)

# output to file
newdata.to_csv('tox21-features_test.csv', index=False)
```

```
print(newdata.shape)
```

The main idea of this provided algorithm is to convert a structure into many numeric attributes to describe the structure in another way. After doing the conversion, I combine the original dataset ‘tox_21’ and the new one ‘tox21-features’ into one ‘tox21_test’ with 12 target attributes and total 124 attributes.

4.2 User interface for input data

Since it’s hard to input a molecule structure, I design the interface only for carcinoma data related to dataset ‘HEPARTWO10k’, but not for ‘tox21’.

4.2.1 Introduction of the interface platform

RuoYi is a fast background development framework based on SpringBoot and Bootstrap. It is a Java EE enterprise-level rapid development platform, based on a combination of classic technologies (Spring Boot, Apache Shiro, MyBatis, Thymeleaf, Bootstrap), some modules built-in such as: department management, role users, menu and button authorization, data permissions, system parameters Log management, notice announcement, etc. There are some main features of it:

Fully responsive layout, which supports all mainstream devices such as computers, tablets, and mobile phones.

Powerful one-key generation function including controller, model, view, menu, etc.

Support multiple data sources, simple configuration can achieve switching.

Support button and data authority can customize department data authority.

Secondary packaging of commonly used js plugins makes the js code concise and easier to maintain.

Perfect XSS prevention and script filtering to completely eliminate XSS attacks.

Maven multi-project dependencies, modules and plug-ins are sub-projects, as loosely coupled as possible, to facilitate module upgrades, increase and decrease modules.

Internationalization supported, server and client supported.

A complete log recording system can be achieved by simple annotation. [15]

4.2.2 Flow of the user interface

First user just need to login by inputting the correct verification code.

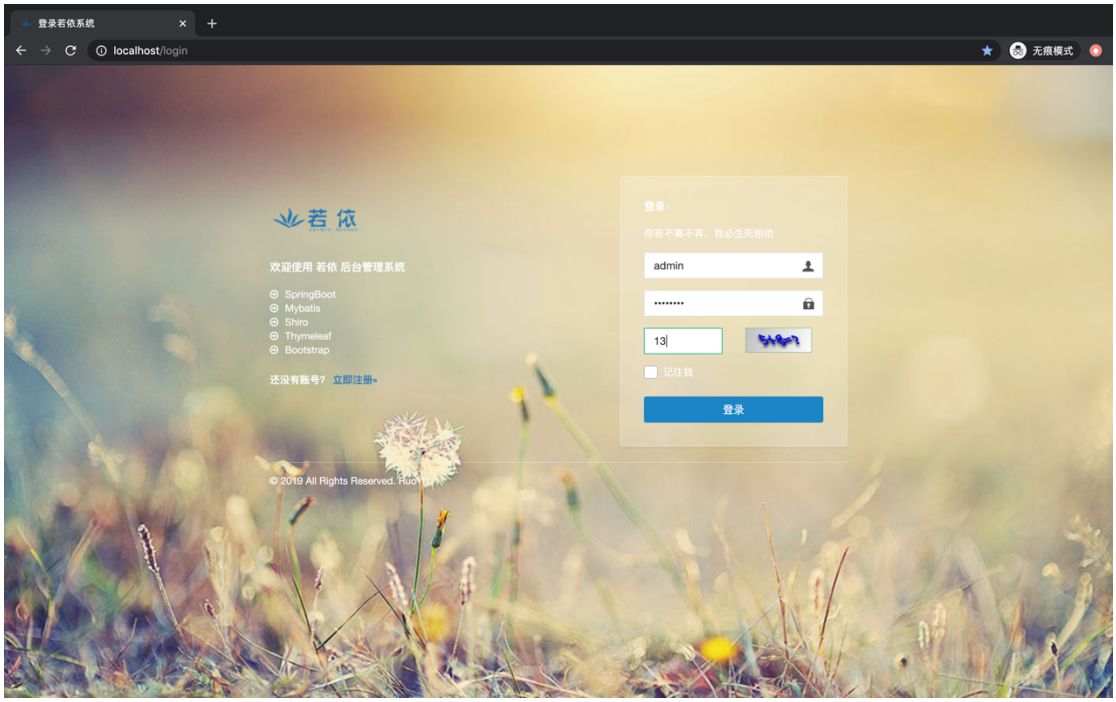

Figure 4-2: Sample 1 of user interface

When user login successfully login, he need to choose ‘carcinoma 检测’ on the right-side bar to open the corresponding interface. Then, click on the ‘+添加’ button, and click to select each value of the data. Once he finishes those data addition operation, click the ‘确认’ button to submit the data input.

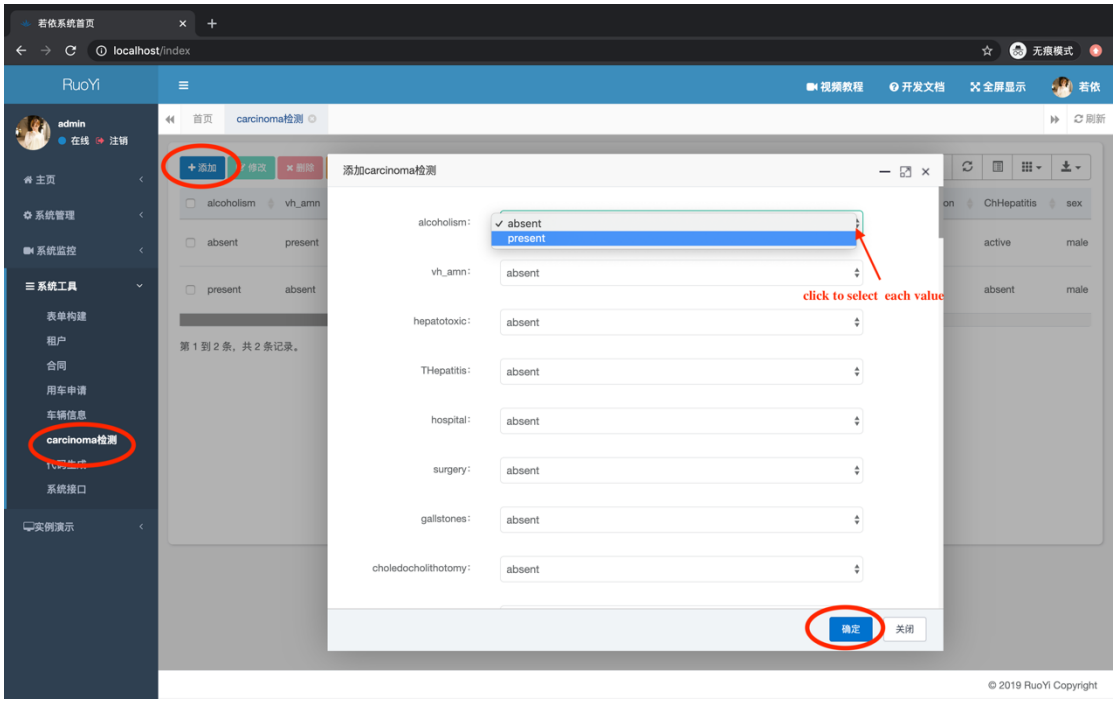

Figure 4-3: Sample 2 of user interface

Figure 4-4 is a sample showing the results after input successfully.

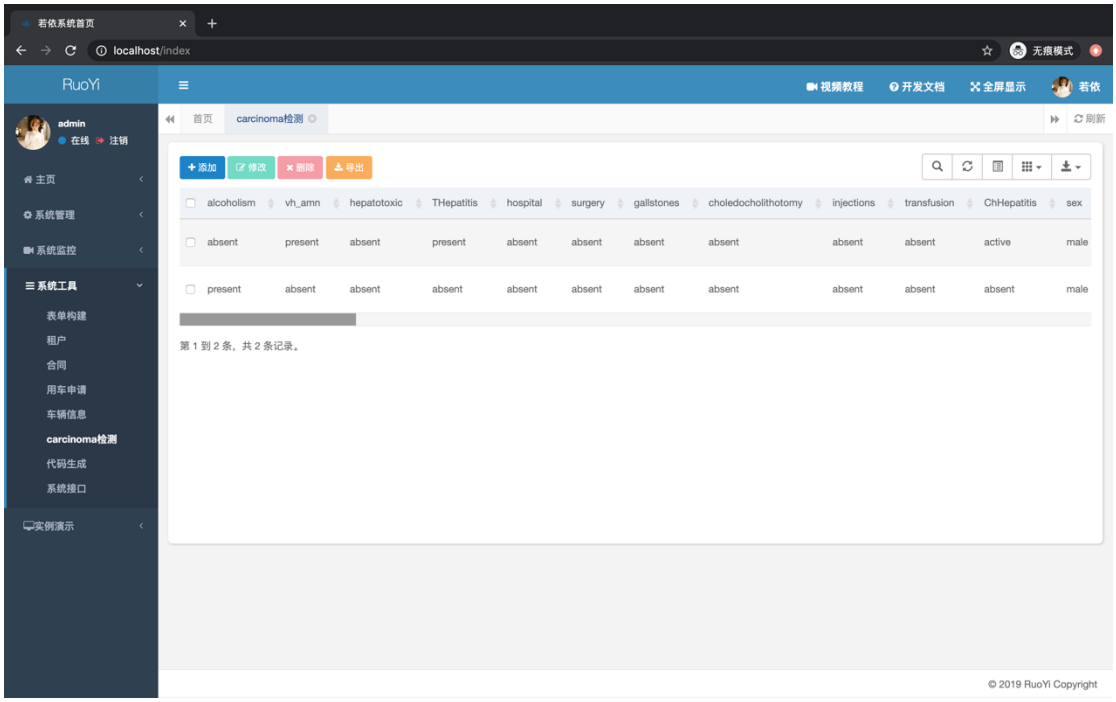

Figure 4-4: Sample 3 of user interface

'Search' function is also provided in the interface. First, user need to click the search button to let it be visible, since the searching content is folded ordinary. Then he need to choose one that he wants to search for, and types in the searching information. By clicking the green '搜索' button, the searching function will affected and show the result.

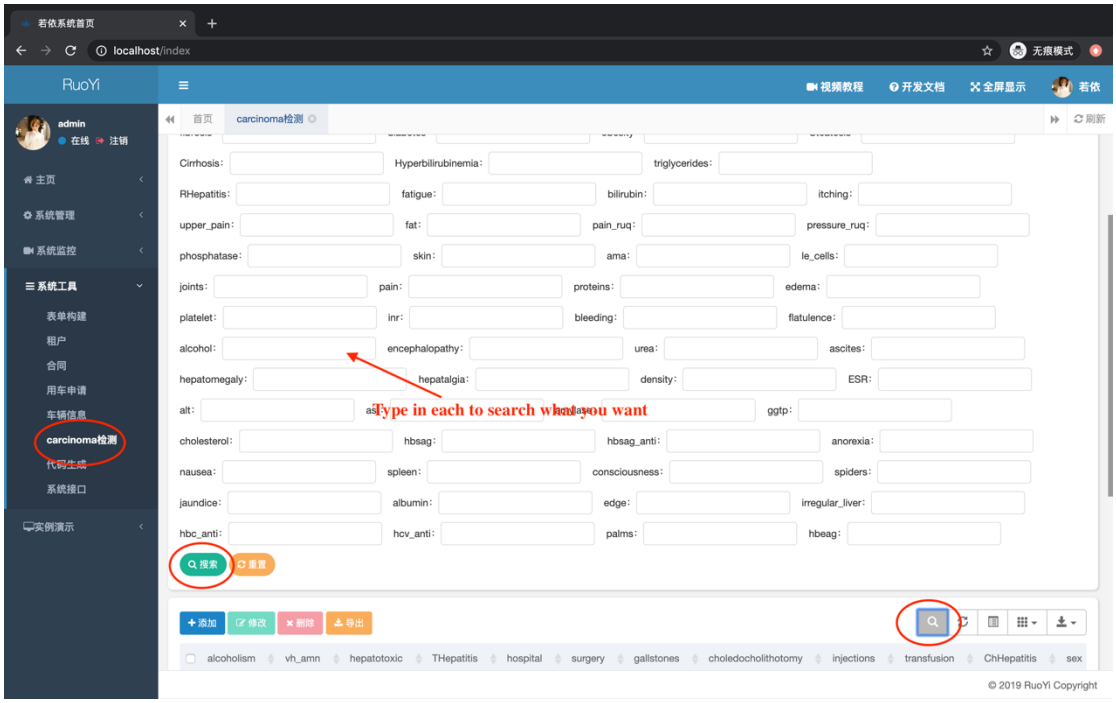

Figure 4-5: Sample 4 of user interface

The following is a step to export the data as a CSV file. Click the circled button and choose ‘CSV’, and the browser will download the csv file.

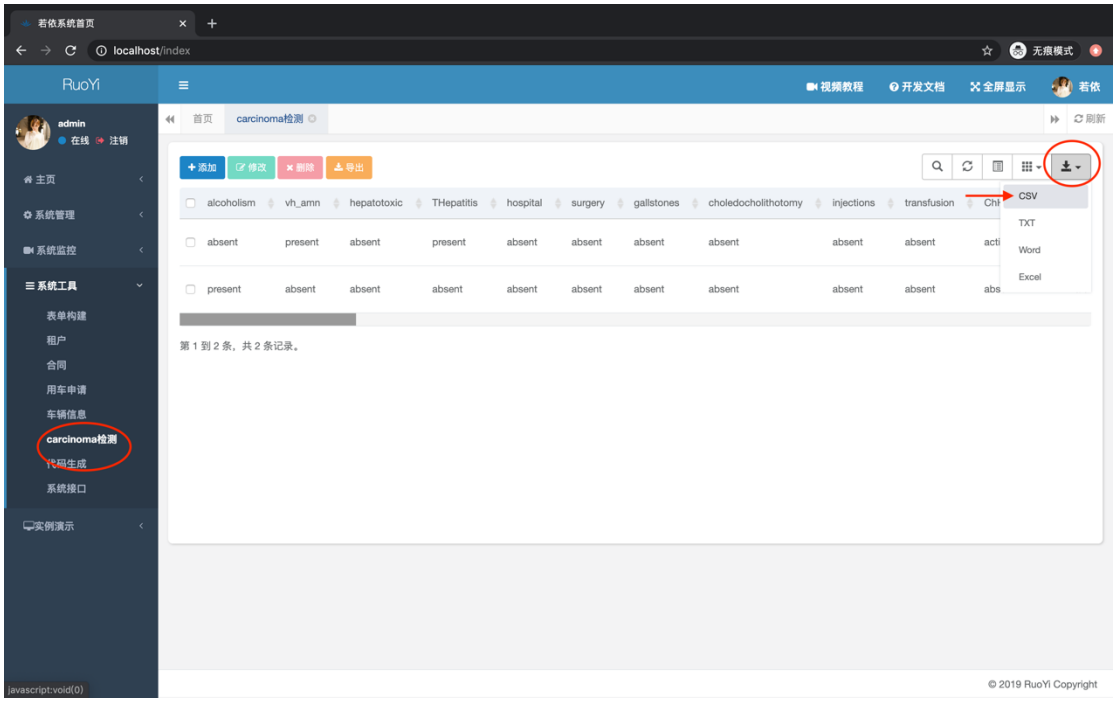

Figure 4-6: Sample 5 of user interface

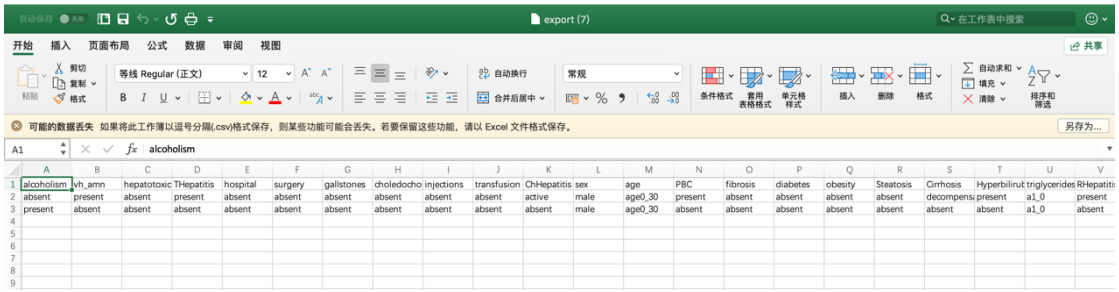

Figure 4-7: Sample result of input exported from user interface

All the user need to do are the above operations, then the CSV file will be download by us to do further machine learning.

4.3 Weka for machine learning

4.3.1 Introduction to Weka

Weka is an open source platform collecting many machine learning algorithms to do data mining. The algorithms provided by Weka can not only be applied directly to a dataset but also called on Java. Weka have many operations for data pre-processing, classification, regression, clustering, association rules, and visualization.

A big advantage of Weka is that it provided plenty of algorithms and the available packages are updated rapidly.

4.3.2 Selected algorithm and results

At the early beginning of the project, we only focused on the dataset 'HEPARTWO10k', we tried different classifiers on it. After comparing the results and learning how to choose an appropriate method, we decided to use FURIA and J48, because our data are mainly in numeric format. It's unlikely to predict by other methods like linear regression, K-means.

classifier											
BayesNet	Summary										
	Correctly Classified Instances	89.30%									
	Incorrectly Classified Instances	10.70%									
	Kappa statistic	0.2145									
	Mean absolute error	0.1388									
	Root mean squared error	0.2913									
	Relative absolute error	116.7577%									
	Root relative squared error	119.56%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.932	0.686	0.953	0.932	0.942	0.217	0.795	0.980	absent	
		0.314	0.068	0.239	0.314	0.271	0.217	0.795	0.206	present	
	Weighted Avg.	0.893	0.647	0.907	0.893	0.9	0.217	0.795	0.931		
	Confusion Matrix										
	a b <-- classified as										
	8731 635 | a = absent										
	435 199 | b = present										
NaiveBayes	Summary										
	Correctly Classified Instances	89.25%									
	Incorrectly Classified Instances	10.75%									
	Kappa statistic	0.2124									
	Mean absolute error	0.1387									
	Root mean squared error	0.2914									
	Relative absolute error	116.732%									
	Root relative squared error	119.5962%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.932	0.688	0.952	0.932	0.942	0.215	0.794	0.980	absent	
		0.312	0.068	0.237	0.312	0.269	0.215	0.794	0.206	present	
	Weighted Avg.	0.893	0.648	0.907	0.893	0.899	0.215	0.794	0.931		
	Confusion Matrix										
	a b <-- classified as										
	8727 639 | a = absent										
	436 198 | b = present										
Logistic	Summary										
	Correctly Classified Instances	93.20%									
	Incorrectly Classified Instances	6.80%									
	Kappa statistic	0.0924									
	Mean absolute error	0.1066									
	Root mean squared error	0.2357									
	Relative absolute error	89.6703%									
	Root relative squared error	96.712%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.99	0.932	0.94	0.99	0.965	0.125	0.793	0.979	absent	
		0.068	0.01	0.326	0.068	0.112	0.125	0.793	0.208	present	
	Weighted Avg.	0.932	0.874	0.901	0.932	0.911	0.125	0.793	0.93		
	Confusion Matrix										
	a b <-- classified as										
	9277 89 | a = absent										
	591 43 | b = present										
RandomForest	Summary										
	Correctly Classified Instances	93.66%									
	Incorrectly Classified Instances	6.34%									
	Kappa statistic	0									
	Mean absolute error	0.1125									
	Root mean squared error	0.2339									
	Relative absolute error	94.6623%									
	Root relative squared error	95.990%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		1.000	1.000	0.937	1.000	0.967	?	0.791	0.980	absent	
		0.000	0.000	?	0.000	?	?	0.791	0.198	present	
	Weighted Avg.	0.937	0.937	?	0.937	?	?	0.791	0.930		
	Confusion Matrix										
	a b <-- classified as										
	9366 0 | a = absent										
	634 0 | b = present										
RandomTree	Summary										
	Correctly Classified Instances	88.94%									
	Incorrectly Classified Instances	11.06%									
	Kappa statistic	0.0835									
	Mean absolute error	0.1124									
	Root mean squared error	0.332									
	Relative absolute error	94.6147%									
	Root relative squared error	136.2514%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.940	0.855	0.942	0.940	0.941	0.084	0.547	0.942	absent	
		0.145	0.060	0.14	0.145	0.143	0.084	0.547	0.076	present	
	Weighted Avg.	0.889	0.805	0.891	0.889	0.89	0.084	0.547	0.887		
	Confusion Matrix										
	a b <-- classified as										
	8802 564 | a = absent										
	542 92 | b = present										
ZeroR	Summary										
	Correctly Classified Instances	93.66%									
	Incorrectly Classified Instances	6.34%									
	Kappa statistic	0									
	Mean absolute error	0.1188									
	Root mean squared error	0.2437									
	Relative absolute error	100.0000%									
	Root relative squared error	100.0000%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		1.000	1.000	0.937	1.000	0.967	?	0.498	0.936	absent	
		0.000	0.000	?	0.000	?	?	0.498	0.063	present	
	Weighted Avg.	0.937	0.937	?	0.937	?	?	0.498	0.881		
	Confusion Matrix										
	a b <-- classified as										
	9366 0 | a = absent										
	634 0 | b = present										
Furia	Summary										
	Correctly Classified Instances	93.52%									
	Incorrectly Classified Instances	6.48%									
	Kappa statistic	0.0338									
	Mean absolute error	0.0709									
	Root mean squared error	0.2512									
	Relative absolute error	59.6828%									
	Root relative squared error	103.0940%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.997	0.978	0.938	0.997	0.966	0.072	0.566	0.945	absent	
		0.022	0.003	0.333	0.022	0.041	0.072	0.566	0.1	present	
	Weighted Avg.	0.935	0.916	0.899	0.935	0.908	0.072	0.566	0.891		
	Confusion Matrix										
	a b <-- classified as										
	9338 28 | a = absent										
	620 14 | b = present										
J48	Summary										
	Correctly Classified Instances	93.08%									
	Incorrectly Classified Instances	6.92%									
	Kappa statistic	0.1215									
	Mean absolute error	0.1093									
	Root mean squared error	0.2516									
	Relative absolute error	91.9500%									
	Root relative squared error	103.2343%									
	Total Number of Instances	10000									
	Detailed Accuracy By Class										
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.988	0.907	0.941	0.988	0.964	0.149	0.563	0.937	absent	
		0.093	0.012	0.335	0.093	0.146	0.149	0.563	0.14	present	
	Weighted Avg.	0.931	0.85	0.903	0.931	0.912	0.149	0.563	0.886		
	Confusion Matrix										
	a b <-- classified as										
	9249 117 | a = absent										
	575 59 | b = present										

Figure 4-8: Results of using different classifiers

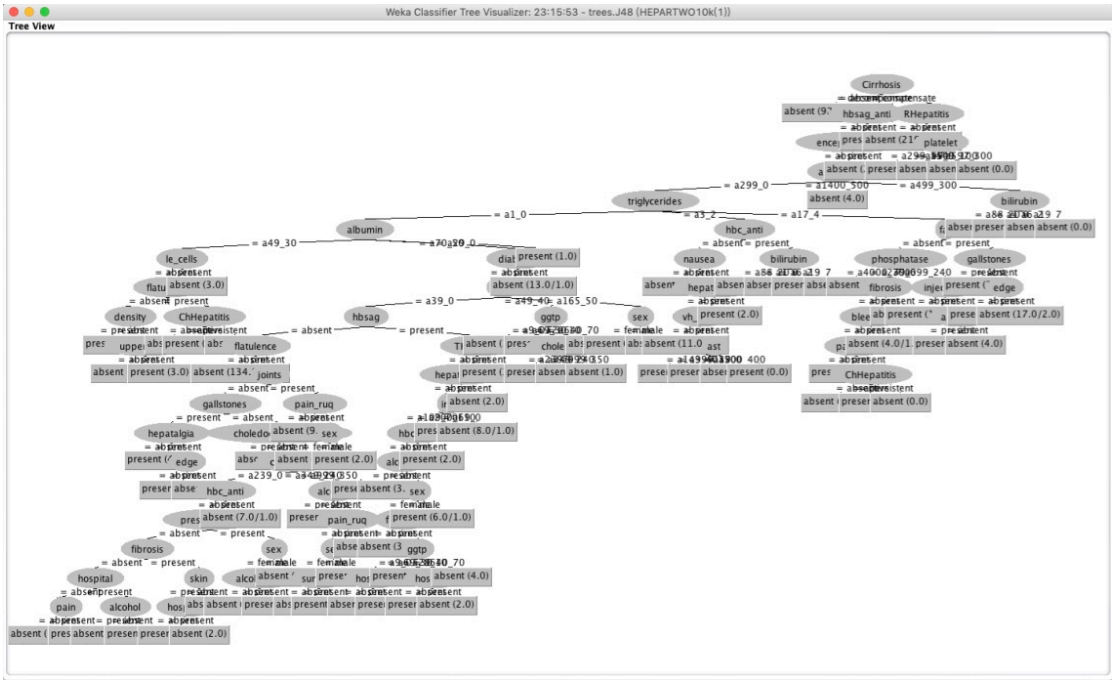

Figure 4-9: Final result of dataset 'HEPARTWO10k' in Weka using J48

From the above tree result, we can know that Cirrhosis is playing an important role in carcinoma result. For every node in the tree, it provide a better understand for the causes of the liver cancer.

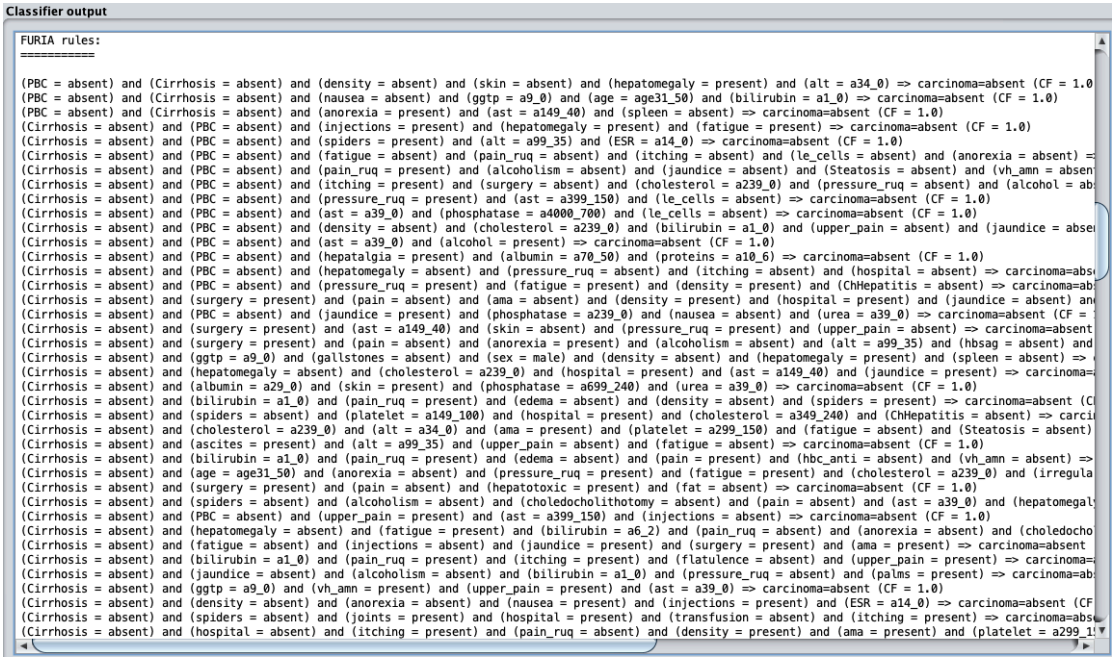

Figure 4-9: Part of result of dataset 'HEPARTWO10k' in Weka using FURIA

The above is the result using FURIA rule algorithm, the visualization will be shown in the other sub-project.

For the dataset 'tox21_test', it should be separated the 12 target values one by one to predict, because it's unable to have multi targets at one time. Therefore, I use RapidMiner to achieve it, which is efficient to have these processes.

4.4 RapidMiner Studio for machine learning

4.4.1 Introduction to RapidMiner

RapidMiner is a world-leading data mining solution, with advanced technology. It covers a wide range of data mining tasks, including various types of data, which can simplify the design and evaluation of the data mining process.

4.4.2 Selected algorithm and results

For datasets 'HEPARTWO10k' and 'tox21_test', I select decision tree in RapidMiner for both. Tree-based learning algorithms are considered to be one of the best and most commonly used supervised learning methods. The tree-based method is able to provide high accuracy, stability, and ability to explain. Unlike linear models, they can map non-linear relationships very well to solve classification or regression questions.

Open in Turbo Prep Auto Model Filter (2 / 2 examples): all

Row No.	prediction(...	confidence(...	confidence(...	alcoholism	vh_amn	hepatotoxic	THepatitis	hospital	surgery	gallstones	choledocho...
1	absent	0.659	0.341	absent	present	absent	present	absent	absent	absent	absent
2	absent	0.956	0.044	present	absent	absent	absent	absent	absent	absent	absent

Figure 4-10: The result of prediction

Figure 4-10 shows the prediction result of new inputted data from user by treating the original data as training ones. The ones in green background is the prediction and the yellow ones are confidence of showing positive and negative results. Besides, the write ones are the retrieved data.

accuracy: 93.35% +/- 0.35% (micro average: 93.35%)

	true absent	true present	class precision
pred. absent	9303	602	93.92%
pred. present	63	32	33.68%
class recall	99.33%	5.05%	

Figure 4-11: The result's performance of dataset 'HEPARTWO10k'

The performance result show the data distribution and an overall accuracy of prediction.

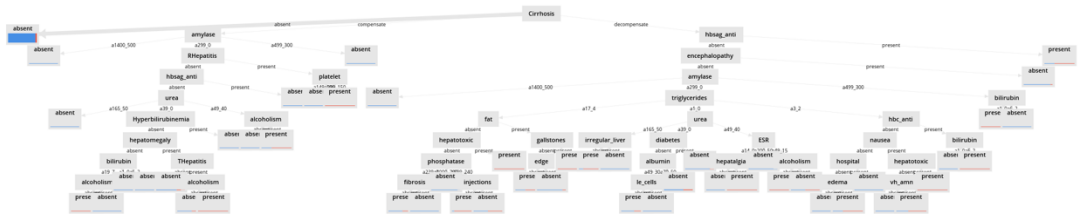

Figure 4-12: The tree result of dataset 'HEPARTWO10k'

In the end, the software will generate a visible decision tree for this specific dataset.

The following figures are results for another dataset ‘tox21_test’.

Table View Plot View

accuracy: 92.50% +/- 0.68% (micro average: 92.50%)

	true 0	true 1	class precision
pred. 0	6185	342	94.76%
pred. 1	166	81	32.79%
class recall	97.39%	19.15%	

Figure 4-13: The result’s performance of dataset ‘tox21_test’

Figure 4-14: Part of the tree result of dataset ‘tox21_test’

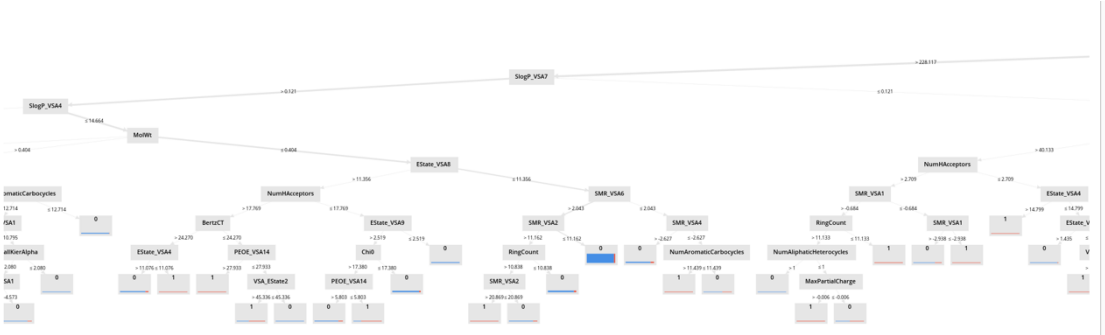

Figure 4-15: Part of the tree result of dataset ‘tox21_test’

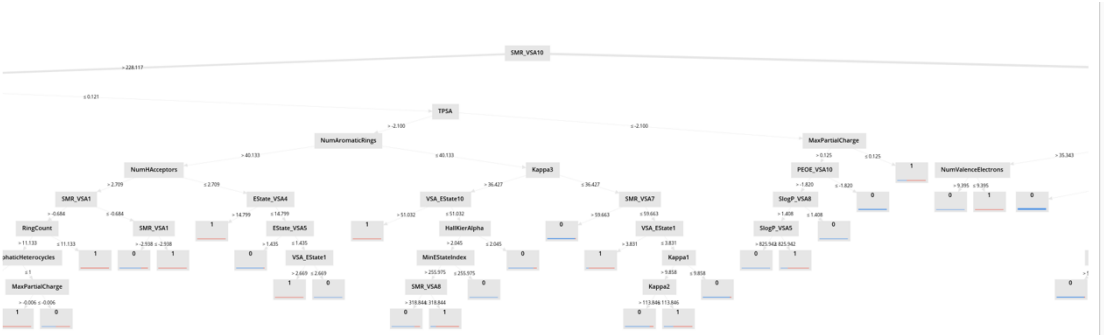

Figure 4-16: Part of the tree result of dataset ‘tox21_test’

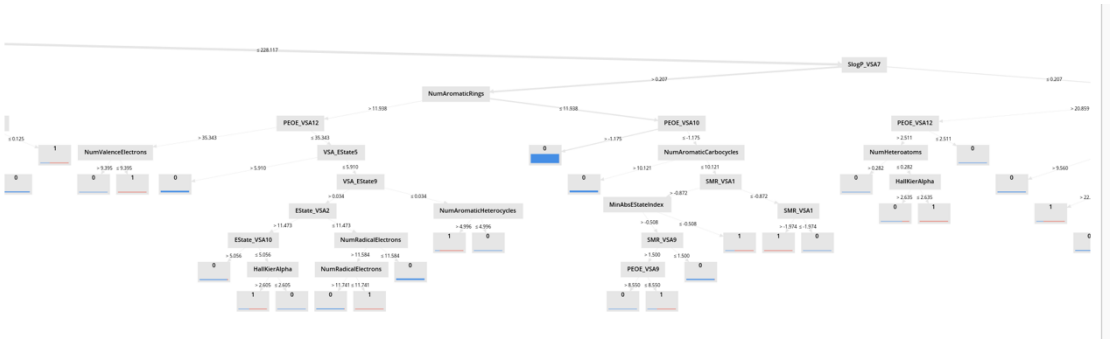

Figure 4-17: Part of the tree result of dataset 'tox21_test'

Figure 4-18: Part of the tree result of dataset 'tox21_test'

CHAPTER 5. Software Design Specificatio

Use Case diagram:

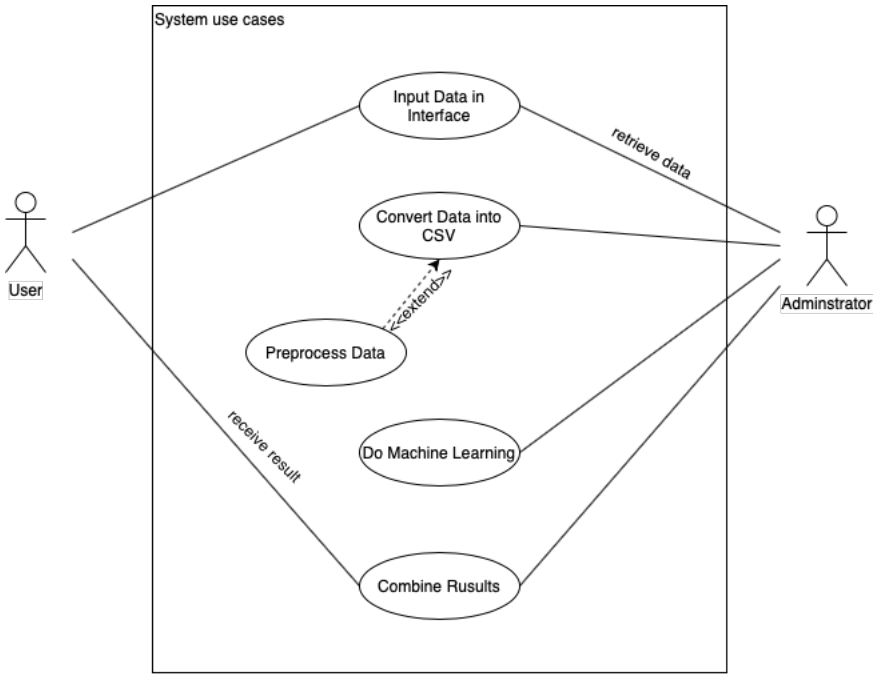

Figure 5-1: Use Case Diagram

Sequence diagram:

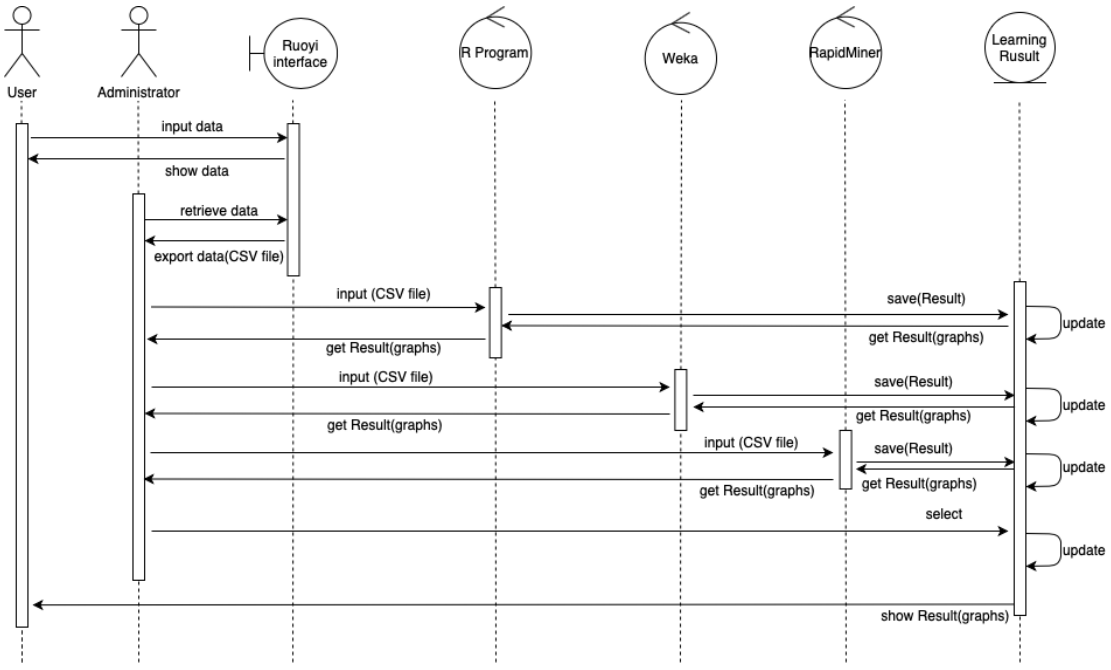

Figure 5-2: Sequence diagram

In this project, we mainly have two actors, user and administrator. For users, they only need to input the data that they want to have prediction. After data input part is

finished, administrator will login again to retrieve data to do machine learning part for users. Then, all machine learning parts are happened in Weka, R program, RapidMiner Studio. Once the results generated, administrator will select the appropriate graph to show and the prediction result.

CHAPTER 6. Implementation Narrative And Description

6.1 Design in RapidMiner

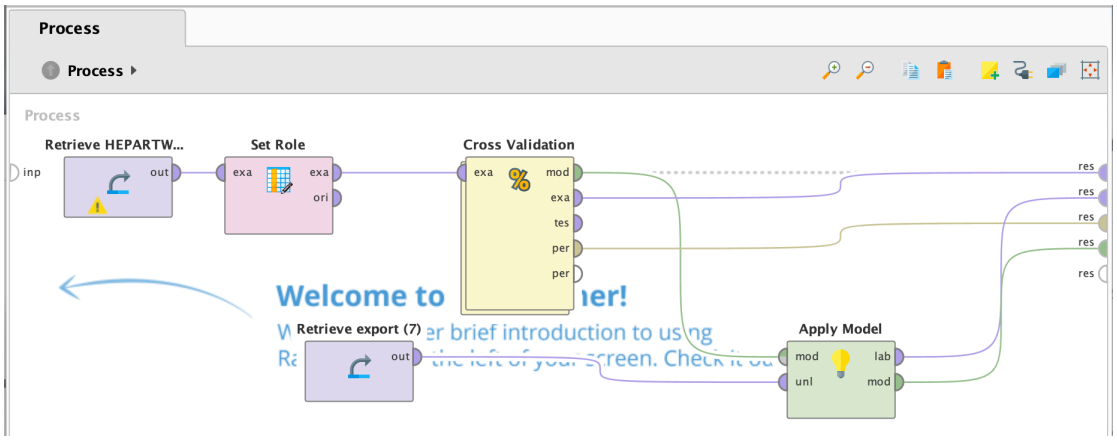

Figure 6-1: Machine learning process of dataset 'HEPARTWO10k'

Firstly, it will retrieve the data, then 'set role' will choose the 'carcinoma' as target value, and finally using decision tree algorithm by cross validation to predict.

In addition, for the CSV file exported from user interface, I firstly set the upwards operation as a model with finished training, then to 'apply model' will automatically apply it to predict the inputted data whether 'carcinoma' present or not.

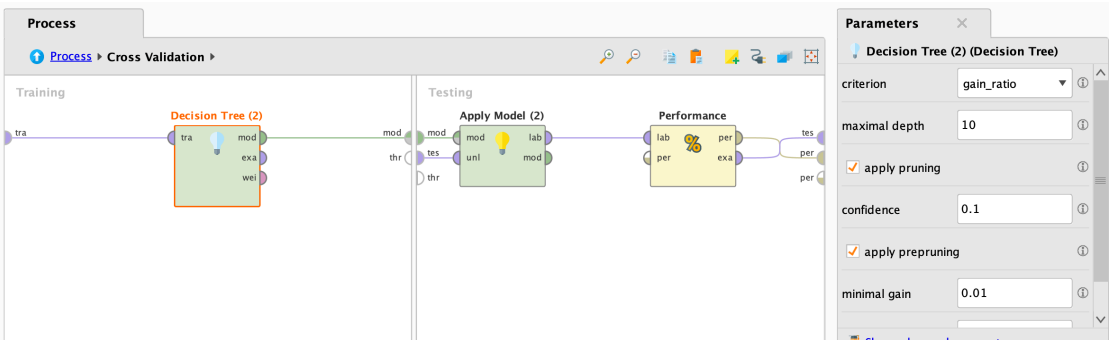

Figure 6-2: Detail cross validation process of dataset 'HEPARTWO10k'

Inside the decision tree, I chose 'gain_ratio' as the criterion, because it showed the best accuracy for this dataset.

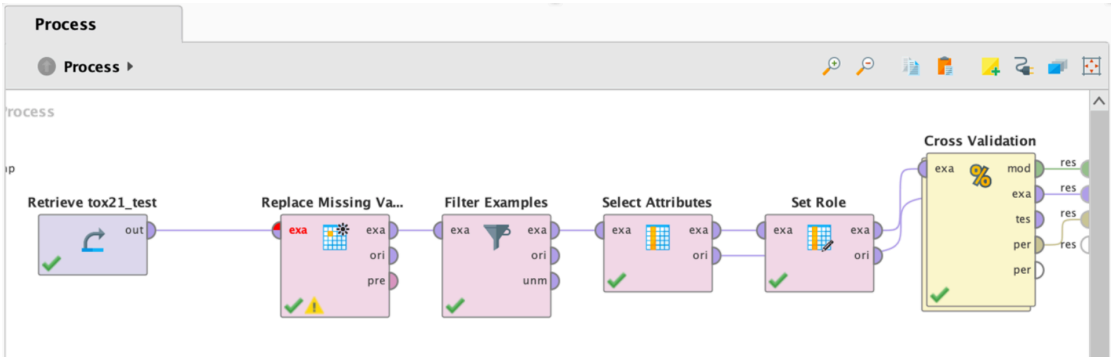

Figure 6-3: Machine learning process of dataset 'tox21_test'

In this process, it will retrieve the data 'tox21_test' first. The 'replace missing value' is to replace some existed infinite data by maximum, 'filter example' will ignore the sequence with missing data for data preprocessing, 'select attribute' is to select the attributes which is the training data, 'set role' is used to decide the target attribute to predict since there are 12 target values in this dataset, the last step is predict using decision tree by cross validation to have a better reliable result.

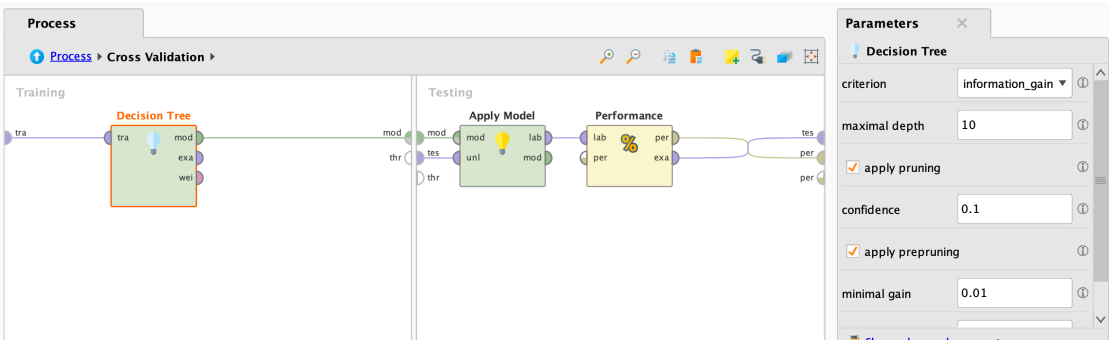

Figure 6-4: Detail cross validation process of dataset 'tox21_test'

Inside the decision tree, I chose 'information gain' as the criterion, since it's more suitable for this dataset to generate a decision tree.

Cross validation in the process is used for avoiding over fitting in decision tree.

6.2 RuoYi user interface

6.2.1 Database setting up

First of all, I create a new table for the carcinoma testing called 'sys_carcinoma', all attributes' names are exactly the same as them in dataset 'HEPARTWO10k', which can be better used for further operations.

```
SET NAMES utf8mb4;
SET FOREIGN_KEY_CHECKS = 0;

DROP TABLE IF EXISTS `sys_carcinoma`;
CREATE TABLE `sys_carcinoma` (
  `carcinoma_id` int(20) NOT NULL AUTO_INCREMENT COMMENT 'id',
  `alcoholism` varchar(30) DEFAULT'',
  `vh_amn` varchar(30) DEFAULT'',
  `hepatotoxic` varchar(30) DEFAULT'',
  `THepatitis` varchar(30) DEFAULT'',
```

```
  `hospital` varchar(30) DEFAULT'',
  `surgery` varchar(30) DEFAULT'',
  `gallstones` varchar(30) DEFAULT'',
  `choledocholithotomy` varchar(30) DEFAULT'',
  `injections` varchar(30) DEFAULT'',
  `transfusion` varchar(30) DEFAULT'',
  `ChHepatitis` varchar(30) DEFAULT'',
  `sex` varchar(30) DEFAULT'',
  `age` varchar(30) DEFAULT '' COMMENT '年龄',
  `PBC` varchar(30) DEFAULT'',
  `fibrosis` varchar(30) DEFAULT'',
  `diabetes` varchar(30) DEFAULT'',
  `obesity` varchar(30) DEFAULT'',
  `Steatosis` varchar(30) DEFAULT'',
  `Cirrhosis` varchar(30) DEFAULT'',
  `Hyperbilirubinemia` varchar(30) DEFAULT'',
  `triglycerides` varchar(30) DEFAULT'',
  `RHepatitis` varchar(30) DEFAULT'',
  `fatigue` varchar(30) DEFAULT'',
  `bilirubin` varchar(30) DEFAULT'',
  `itching` varchar(30) DEFAULT'',
  `upper_pain` varchar(30) DEFAULT'',
  `fat` varchar(30) DEFAULT'',
  `pain_ruq` varchar(30) DEFAULT'',
  `pressure_ruq` varchar(30) DEFAULT'',
  `phosphatase` varchar(30) DEFAULT'',
  `skin` varchar(30) DEFAULT'',
  `ama` varchar(30) DEFAULT'',
  `le_cells` varchar(30) DEFAULT'',
  `joints` varchar(30) DEFAULT'',
  `pain` varchar(30) DEFAULT'',
  `proteins` varchar(30) DEFAULT'',
  `edema` varchar(30) DEFAULT'',
  `platelet` varchar(30) DEFAULT'',
  `inr` varchar(30) DEFAULT'',
  `bleeding` varchar(30) DEFAULT'',
  `flatulence` varchar(30) DEFAULT'',
  `alcohol` varchar(30) DEFAULT'',
  `encephalopathy` varchar(30) DEFAULT'',
  `urea` varchar(30) DEFAULT'',
  `ascites` varchar(30) DEFAULT'',
  `hepatomegaly` varchar(30) DEFAULT'',
  `hepatalgia` varchar(30) DEFAULT'',
  `density` varchar(30) DEFAULT'',
  `ESR` varchar(30) DEFAULT'',
  `alt` varchar(30) DEFAULT'',
  `ast` varchar(30) DEFAULT'',
  `amylase` varchar(30) DEFAULT'',
  `ggtp` varchar(30) DEFAULT'',
  `cholesterol` varchar(30) DEFAULT'',
  `hbsag` varchar(30) DEFAULT'',
  `hbsag_anti` varchar(30) DEFAULT'',
  `anorexia` varchar(30) DEFAULT'',
  `nausea` varchar(30) DEFAULT'',
  `spleen` varchar(30) DEFAULT'',
  `consciousness` varchar(30) DEFAULT'',
  `spiders` varchar(30) DEFAULT'',
  `jaundice` varchar(30) DEFAULT'',
  `albumin` varchar(30) DEFAULT'',
  `edge` varchar(30) DEFAULT'',
  `irregular_liver` varchar(30) DEFAULT'',
  `hbc_anti` varchar(30) DEFAULT'',
  `hcv_anti` varchar(30) DEFAULT'',
  `palms` varchar(30) DEFAULT'',
  `hbeag` varchar(30) DEFAULT'',


  PRIMARY KEY (`carcinoma_id`)
) ENGINE=InnoDB AUTO_INCREMENT=200 DEFAULT CHARSET=utf8 COMMENT='carcinoma 检测
表';
```

After being created successfully, we can see from the Navicat Premium

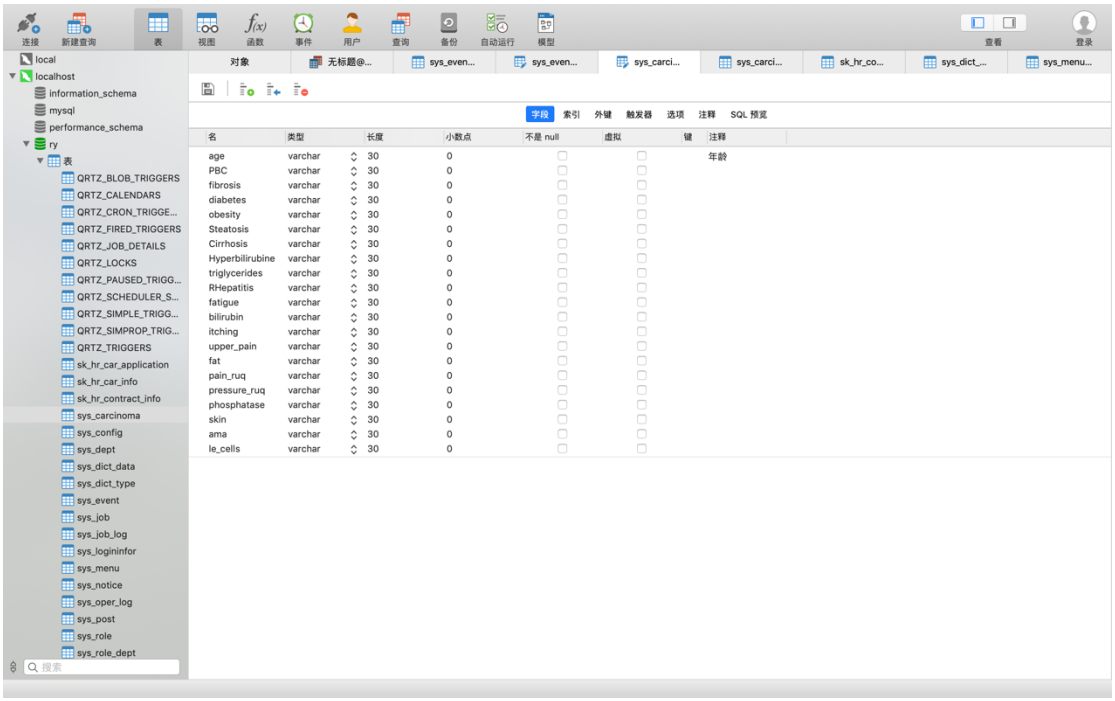

Figure 6-5: Created table of the database shown on Navicat Premium

Besides, the following code is to insert a new function 'carcinoma test' into the menu of the interface. It's important to have the proper 'parent_id', which makes it in the correct position of the menu in the interface.

```
-- 菜单 SQL
insert into sys_menu (menu_name, parent_id, order_num, url,menu_type, visible,
perms, icon, create_by, create_time, update_by, update_time, remark)
values('carcinoma 检 测 ', '3', '1', '/system/carcinoma', 'C', '0',
'system:carcinoma:view', '#', 'admin', '2018-03-01', 'ry', '2018-03-01',
'carcinoma 检测菜单');

-- 按钮父菜单 ID
SELECT @parentId := LAST_INSERT_ID();

-- 按钮 SQL
insert into sys_menu  (menu_name, parent_id, order_num, url,menu_type, visible,
perms, icon, create_by, create_time, update_by, update_time, remark)
values('检测查询', @parentId, '1', '#', 'F', '0', 'system:carcinoma:list',
'#', 'admin', '2018-03-01', 'ry', '2018-03-01', '');

insert into sys_menu  (menu_name, parent_id, order_num, url,menu_type, visible,
perms, icon, create_by, create_time, update_by, update_time, remark)
values('检测新增', @parentId, '2', '#', 'F', '0', 'system:carcinoma:add',
'#', 'admin', '2018-03-01', 'ry', '2018-03-01', '');

insert into sys_menu  (menu_name, parent_id, order_num, url,menu_type, visible,
perms, icon, create_by, create_time, update_by, update_time, remark)
values('检测修改', @parentId, '3', '#', 'F', '0', 'system:carcinoma:edit',
'#', 'admin', '2018-03-01', 'ry', '2018-03-01', '');

insert into sys_menu  (menu_name, parent_id, order_num, url,menu_type, visible,
perms, icon, create_by, create_time, update_by, update_time, remark)
values('检测删除', @parentId, '4', '#', 'F', '0', 'system:carcinoma:remove',
'#', 'admin', '2018-03-01', 'ry', '2018-03-01', '');
```

6.2.2Algorithm of ‘export’ function

6.2.2.1

The front-end will use a encapsulated method ‘$.table.init’, transfer to ‘exortUrl’.

Part of the code is shown:

```
var options = {
                url: prefix + "/list",
                createUrl: prefix + "/add",
                updateUrl: prefix + "/edit/{id}",
                removeUrl: prefix + "/remove",
                exportUrl: prefix + "/export",
                modalName: "carcinoma 检测",
                showExport: true,
                columns: [{
                        checkbox: true
                    },
                        {
                                field : 'carcinomaId',
                                title : 'id',
                                visible: false
                        },
                        {
                                field : 'alcoholism',
                                title : 'alcoholism',
                                sortable: true
                        }]
            };
            $.table.init(options);
```

6.2.2.2

Add an export button

Example of the code:

```
<a        class="btn        btn-warning"        onclick="$.table.exportExcel()"
shiro:hasPermission="system:carcinoma:export">
    <i class="fa fa-download"></i> 导出
</a>
```

6.2.2.3

Add ‘@Excel’ annotation on entity variables

Example of the code:

```
    private Integer carcinomaId;

    @Excel(name = "alcoholism")
    private String alcoholism;

    @Excel(name = "vh_amn")
    private String vh_amn;

    @Excel(name = "hepatotoxic")
    private String hepatotoxic;
```

6.2.2.4

Add export method in Controller

Example of the code:

```
/**
* 导出 carcinoma 检测列表
*/
@RequiresPermissions("system:carcinoma:export")
@PostMapping("/export")
@ResponseBody
public AjaxResult export(Carcinoma carcinoma)
{
    List<Carcinoma> list = carcinomaService.selectCarcinomaList(carcinoma);
     ExcelUtil<Carcinoma> util = new ExcelUtil<Carcinoma>(Carcinoma.class);
     return util.exportExcel(list, "carcinoma");
}
```

6.2.3 A small bug existed before and solution

Name of each attribute in HTML file must be exactly the same as JSON

Part of the HTML file:

```
{
    field : 'alcoholism',
    title : 'alcoholism',
    sortable: true},
```

JSON:


```
{
    "total": 1,
    "rows": [{
            "searchValue": null,
            "createBy": null,
            "createTime": null,
            "updateBy": null,
            "updateTime": null,
            "remark": null,
            "params": {},
            "carcinomaId": 209,
            "alcoholism": "present",
            "vh_amn": "absent",
            "hepatotoxic": "absent",
            "hospital": "absent",
            "surgery": "absent",
            "gallstones": "absent",
            "choledocholithotomy": "absent",
            "injections": "absent",
            "transfusion": "absent",
            "sex": "male",
            "age": "age0_30",
            "fibrosis": "absent",
            "diabetes": "absent",
            "obesity": "absent",
            "triglycerides": "a1_0",
            "fatigue": "absent",
            "bilirubin": "a1_0",
            "itching": "absent",
            "upper_pain": "absent",
            "fat": "absent",
            "pain_ruq": "absent",
            "pressure_ruq": "absent",
            "phosphatase": "a239_0",
            "skin": "absent",
            "ama": "absent",
            "le_cells": "absent",
            "joints": "absent",
            "pain": "absent",
            "proteins": "a5_2",
            "edema": "absent",
            "platelet": "a99_0",
            "inr": "a69_0",
            "bleeding": "absent",
            "flatulence": "absent",
```

```
            "alcohol": "absent",
            "encephalopathy": "absent",
            "urea": "a39_0",
            "ascites": "absent",
            "hepatomegaly": "absent",
            "hepatalgia": "absent",
            "density": "absent",
            "alt": "a34_0",
            "ast": "a39_0",
            "amylase": "a299_0",
            "ggtp": "a9_0",
            "cholesterol": "a239_0",
            "hbsag": "absent",
            "hbsag_anti": "absent",
            "anorexia": "absent",
            "nausea": "absent",
            "spleen": "absent",
            "consciousness": "absent",
            "spiders": "absent",
            "jaundice": "absent",
            "albumin": "a29_0",
            "edge": "absent",
            "irregular_liver": "absent",
            "hbc_anti": "absent",
            "hcv_anti": "absent",
            "palms": "absent",
            "hbeag": "absent",
            "thepatitis": "absent",
            "chHepatitis": "absent",
            "pbc": "absent",
            "steatosis": "absent",
            "cirrhosis": "absent",
            "hyperbilirubinemia": "absent",
            "rhepatitis": "absent",
            "esr": "a14_0"
        }],
        "code": 0
    }
```


During the user interface coding, it happened to me once that the field name is in capital, which is different to the JSON one. That lead to a bug that could not show the input from user in the interface. Therefore, we should pay attention to this small but important programming matter.

CHAPTER 7. System Quality

As you can see from Figure 4-11 and Figure 4-13, the accuracies are above 90 percent. Besides, from figure 4-10, confidences of the predictions of input data are also over 65 percent. These high percentage of the accuracy shows the system could be in good quality to trust for.

For some datasets related to medical like the 'HEPARTWO10k' in this project, I think it is a reliable system to help users like doctor to better judge on the patients' conditions. On the other hand, the tree result show cirrhosis is a significant symbol for liver cancer, which can help doctor to decide some screening objects to specific disease. To achieve a practical medical use for the society is our goal from the beginning, and doing this project can at least bring us to the right road and carry on.

CHAPTER 8. Ethics And Professional Conduct

Security of data attracts much attention nowadays, we also noticed about this topic while we decided to set up a user interface for inputting. Therefore, I filtered so much platform until I chose the one 'RuoYi'. It is built based on the Apache Shiro framework, a security framework for Java. Shiro can help complete operations: authentication, authorization, encryption, session management, integration with the Web, caching, etc. The security framework make sure we are able to protect security of data if we make the website public.

All codes are design by ourselves or from open source, there is no any illegal copying from others. We show respect to all others' contributions.

All licenses of software used in this my part of project are authorized:

RStudio, Version 1.2.5019, free;

Navicat Premium, Version 12.1.23, Licensed to LIHENG LIANG;

Weka, Version 3.8.3, free;

RapidMiner Studio, Version 9.6, Educational Edition to Simon Fong;

IntelliJ IDEA, 2019.1.3 (Ultimate Edition), Licensed to LIHENG LIANG;

CHAPTER 9. Summary

After finished this project with my teammates and under careful guidance from my supervisors, I am so delighted to have such a chance to enhance myself in programming skills. I would like to be grateful again to the cooperation with my teammates and the great detailed suggestion from supervisors Shirley Siu and Simon Fong. I learned how to schedule well for one complete project and how to communicate well with teammates.

There is an interesting case during the project that I learned from an online video about RapidMiner issued on 2013. One day, I found out some question in the hands-on using, then I tried to contact the instructor of that course. Fortunately, I can still reach the constructor for the contact information after such a long time. I receive so much help from this enthusiastic constructor. Then, I started to understand the most important thing to overcome difficulties is communication.

In addition, there are also some shortages in this project. Therefore, we have some future expectation to the project. For example, in this period, we still cannot find an appropriate way to combine different machine learning result like we can generate many from various algorithm into one. We hope that we can work out a method to combine them and optimize. Moreover, the user interface is separated from the machine learning, we want to package them into one, so that once the user finishes inputting, the machine learning operation will start automatically.

All in all, it’s a wonderful experience for me to have this project especially in this tough period. Hope this is a start for me to contribute myself for a everyone’s better living environment by programming, and to bring more practical improvement.

CHAPTER 10. Refferences

[1] D. Curiac, G. Vasile, O. Banias, C. Volosencu and A. Albu, "Bayesian network model for diagnosis of psychiatric diseases," 2009. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5196055

[2] Ryan Roberts, “Machine Learning: The Ultimate Beginners Guide For Neural Networks, Algorithms, Random Forests and Decision Trees Made Simple”.[Online]. Available: http://www.doc88.com/p-3532862511967.html

[3] C. Anthony, F. Norman, M. William and R. Lukasz, "From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support," February 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336571600004X

[4] R. Nagarajan, M. Scutari and S. Lèbre, *Bayesian Networks in R with Applications in Systems Biology*, 2013.

[5] M. Scutari and J.-B. Denis, *Bayesian Networks with Examples in R*, 2014.

[6] U. Laura, "Advantages and challenges of Bayesian networks in environmental modelling," May 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304380006006089

[7] A. Peter, F. Geert, T. Dirk, M. Yves, M. Bart De, "Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection," 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365703000538

[8] bnlearn, "bnlearn - an R package for Bayesian network learning and inference". [Online]. Available: http://bnlearn.com/

[9] José A. GámezJuan L. MateoEmail authorJosé M. Puerta, "Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighbourhood, " 2001. [Online]. Available: https://link.springer.com/article/10.1007/s10618-010-0178-6

[10] BayesiaLab, "The Leading Desktop Software for Bayesian Networks". [Online]. Available: https://www.bayesia.com/

[11] R, "The R Project for Statistical Computing". [Online]. Available: https://www.r-project.org/

[12] Mayur Kulkarni, ‘Decision Trees for Classification: A Machine Learning Algorithm’.September 7, 2017. [Online]. Available: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html

[13] chamie, ‘决策树 J48 算法’. [Online]. Available: https://www.cnblogs.com/chamie/p/4523976.html

[14] Jens Hühn, Eyke Hüllermeier, ‘FURIA: An Algorithm For Unordered Fuzzy Rule Induction’. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.447.2303&rep=rep1&type=pdf

[15] [Online]. Available: http://doc.ruoyi.vip/ruoyi/document/kslj.html#项目简介

CHAPTER 11. Appendix

Software and environment setting needed for our project:

1) Computer with Windows or MacOS
2) JDK >= 1.8
3) Mysql >= 5.5.0
4) Maven >= 3.0
5) Navicat Premium
6) RStudio
7) RapidMiner Studio
8) Weka