Research Article

It Medical Team

www.itmedicalteam.pl

Journal of Biomedical Sciences

ISSN 2254-609X

2023

Vol.12 No.1:92

Using Two Methods of Decision Tree and Regression in Identifying the Most Effective Factors in the Occurrence of Hepatitis C Diseases, Hepatitis C with Fibrosis and Hepatitis C with Cirrhosis and Checking the Results and Accuracy of These Two Methods

Arefe Bagheri[1*], and Mina Kalini[2]

[1]Department of Medical Sciences, Al-Zahra University of Mashhad, Tehran, Iran

[2]Department of Medical Sciences, Ashrafi University of Esfahani, Isfahan Province, Iran

[*]**Corresponding author:** Arefe Bagheri, Department of Medical Sciences, Al-Zahra University of Mashhad, Tehran, Iran, Tel: 00989130157962; E-mail: arefeh.br@gmail.com

Received date: August 22, 2022, Manuscript No. IPJBS-22-12960; **Editor assigned date:** August 24, 2022, PreQC No. IPJBS-22-12960 (PQ); **Reviewed date:** September 08, 2022, QC No. IPJBS-22-12960; **Revised date:** January 20, 2023, Manuscript No. IPJBS-22-12960 (R); **Published date:** January 27, 2023, DOI: 10.36648/2254-609X.12.1.92

Citation: Bagheri A, Kalini M (2023) Using Two Methods of Decision Tree and Regression in Identifying the Most Effective Factors in the Occurrence of Hepatitis C Diseases, Hepatitis C with Fibrosis and Hepatitis C with Cirrhosis and Checking the Results and Accuracy of These Two Methods. J Biomed Sci Vol:12 No:1

Abstract

One of the common disorders of the liver is the hepatitis C virus, which after a period causes serious damage to the liver and the failure of this important organ. This disease can improve or progress to liver fibrosis or cirrhosis. In this regard, this research focuses on the identification of effective factors and the extent of these factors in the occurrence of hepatitis C disease and two other types along with liver fibrosis or cirrhosis so that doctors and patients can pay more attention to these factors and avoid the occurrence of the disease. In this research, using the information related to blood donors, the implementation operation is carried out to identify the effective factors in the occurrence of disease in these persons. This research uses the two techniques of decision tree and regression to carry out the implementation work, and finally the efficiency of each is examined and compared. At the end of this research and after examining the results of modeling methods, the three factors ALB, AST and CHE are known as the most effective factors in diagnosing the occurrence of hepatitis C and its advanced types. The results of this research show that the accuracy of diagnosis in this research with the decision tree method equals 94.57% and with the regression technique has an accuracy of 91.28%.

Keywords: Hepatitis C; Decision tree; Regression; Data mining; Liver diseases

Introduction

Hepatitis is inflammation of the liver. 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 [1].

Types of hepatitis include A, B, C, D, and E. These 5 types are of the greatest concern due to the burden of disease and death caused by them and the potential of their spread. In particular, types B and C lead to chronic disease in hundreds of millions of people and are the most common cause of liver cirrhosis and cancer.

Hepatitis A and E are usually caused by consuming contaminated food or water. Hepatitis B, C, and D usually occur as a result of periodical contact with contaminated body fluids. Common methods of transmission of these viruses include receiving contaminated blood or blood products, invasive medical procedures using contaminated equipment, the transmission of hepatitis B from mother to infant at birth, from family members to the child, and sexual contact.

Types of hepatitis A, B, and C are diagnosed by symptoms, physical examination, and blood tests. Sometimes imaging studies such as ultrasound or CAT scan and liver biopsy are also used.

Hepatitis C is most commonly transmitted through contact with infected blood and is usually a long-term infection with no symptoms. Hepatitis C can lead to scarring of the liver or cirrhosis. It should be noted that there is no vaccine to prevent it [2]. The ways of its transmission are:

- Sharing of dirty syringes.
- Direct contact with infected blood or body fluids of an infected person.
- Blood transfusion by an infected person.
- Sexual contact with someone infected.

Materials and Methods

Due to the prevalence and dangerous damage of this disease, the prediction of hepatitis C infection has been prioritized for doctors and scientists, but until now, the prediction of events related to hepatitis C in clinical practice has not been able to achieve high accuracy. In this context, electronic health records (also called medical records) can be considered as a useful

source of information in order to reveal hidden and unclear relationships between patients' data, not only for research but also for clinical procedures. For this purpose, several screening studies have been carried out in the past years, which include different conditions and populations and with different data sources, to deepen the knowledge of risk factors. Today, hepatitis C prevention modeling is also in terms of achieving high predictive accuracy [3]. And identifying the driving factors is still a problem. Most of the models developed for this purpose with limited interpretation of the predictor variables are obtained with only moderate accuracy. Newer models show improvements. Although scientists have identified a wide range of predictors and indicators, there is no common consensus regarding their relative influence on the prediction of this disease. This situation is mainly due to the lack of reproducibility, which prevents definitive conclusions about the importance of the identified factors. Furthermore, this lack of reproducibility strongly affects model performance (generalization to external validation datasets is often inconsistent and achieves only moderate discrimination). As a result, the risk scores distilled from the models face similar problems and limit their reliability. Such uncertainty has led to the proliferation of new risk scores that appear in articles with different results in the past years.

Machine learning, which is specifically applied to medical records, can be an effective tool for predicting hepatitis C in any patient with symptoms of this disease, and it can also identify the most important clinical features (or risk factors) that It may lead to this disease, to diagnose. Scientists can use machine learning not only for clinical prediction, but also for feature ranking [4]. Computational intelligence, especially when applied to medical records, along with imaging, shows its predictive power.

In addition, applied deep learning studies and meta-analysis in this field have also recently appeared in the literature, which improve the performance of human experts, albeit with lower accuracy (0.59 *vs.* 0.75). Today, with the help of data mining methods, the relationship between different effective or ineffective factors in a subject is determined, and given that data mining is a useful tool in extracting knowledge from mass data, which shows the hidden connections between them [5]. However, the medical community has also turned to these techniques.

Data mining is not limited to the use of technologies and will use anything that is effective for it. Nevertheless, statistics and computer are the most used sciences and technologies used in data mining.

Basics of data mining

A set of methods that can be applied to large and complex databases in order to discover hidden and interesting patterns hidden among the data is called data mining [6]. And in other words, the process of extracting and discovering correlations and useful patterns from a large amount of raw data, which is done using algorithms and intelligent mechanisms, is called data mining. Data mining methods are almost always computationally expensive.

Main applications of data mining

Among other applications of data mining are:

- Discovering patterns among data
- Quantitative prediction of results
- Obtaining practical information
- Focus on big data

In general, the data mining process, in addition to helping to remove irrelevant and useless data from the data set, on the other hand, provides very useful and practical information to organizations and also speeds up decision-making processes [7].

Results and Discussion

The process of doing data mining

As you can see in Figure 1, data mining is generally performed in 6 main stages, at first, the required data is collected and processed and cleaned, that is, the extra data is removed and only the required data is entered into the system. In the next step, the pattern between the data will be discovered and evaluated, and then the algorithm and data mining methods will be performed on the data [8].

Finally, the information obtained from the data mining process is presented in the form of human-understandable formats such as graphs, images, reports, etc., and the desired knowledge extracted from the mass of raw data will be provided to the organization.

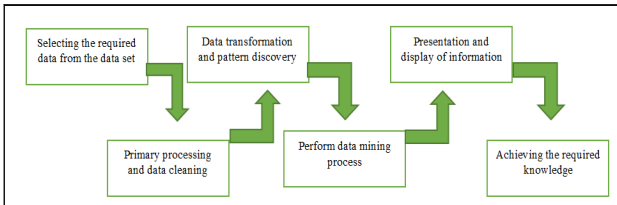

Figure 1: Data mining process.

Introduction of hepatitis C

Hepatitis C is one of the most important causes of liver diseases in the world. Hepatitis C seems to be the main cause of liver diseases and death due to liver diseases in the coming years. About 71 million people around the world are suffering from chronic hepatitis C; of course, many of them are unaware of their infection [9]. The prevalence of this disease is average in people with thalassemia (16.6%), hemophilia (54%), people undergoing dialysis (8.3 percent), and injection drug addicts (51.4 percent).

Clinical care for hepatitis C patients has grown very well due to the progress in understanding the pathophysiology of the disease and due to the development of diagnostic-therapeutic and preventive methods, and eradication has been considered by the World Health Organization by 2030. To eradicate the disease, the need is to screen and treat patients. ELISA test is used for initial screening and the presence of antibodies against hepatitis C is checked. Rapid Disease Diagnosis Tests (RDT) and antibody testing in serum, plasma, fingertip capillary blood, whole blood, and oral saliva can also be used for primary

screening. If the screening test is positive, it is necessary to confirm the diagnosis with PCR or HCV core Ag [10]. Although the central antigen has less sensitivity than PCR, it has gained great importance due to its lower cost and relatively good sensitivity. A positive ELISA test and a negative PCR can be due to the following reasons: 1) False positive, 2) The disease has recovered spontaneously, 3) The patient has been treated, and 4) low levels of the virus in the blood that could not be detected.

It should be noted that despite proper treatment and eradication of the disease in the affected person, Elisa may remain positive until the end of her life. Therefore, PCR and HCV core Ag are used to follow up on patients. Hepatitis C has different genotypes (1 to 7) and currently, due to its genetic diversity, it is not possible to make a vaccine. Screening in high-risk groups is preferable to screening in the community [11]. In injecting drug addicts, annual screening is recommended. In the past, treatment with interferon and ribavirin was for 24 to 48 weeks. In addition to low response and many side effects, this treatment also costs a lot. In 2011, protease inhibitors, first-generation DAAs (telaprevir-boceprevir) were added as a third drug to the previous treatment. These drugs also had many side effects such as rash, anemia, and many drug interactions. In 2013, a new drug called sofosbuvir was proposed for the treatment of hepatitis C, which shortened the treatment period and had fewer side effects compared to the third treatment based on protease inhibitors.

The 2018 EASL recommendations on the treatment of hepatitis C also suggest other DAAs such as glicaprovir, pibrentasvir, grazoprevir, and albatasvir for treatment. Also, treatments of 8 to 16 weeks and 28 weeks are given in special cases. Treatments without sofosvir are also considered in new treatments. Before treatment, it is better to determine quantitative PCR and virus genotype. If there is a problem in doing the above, the existence of a positive qualitative PCR without determining the genotype is enough for treatment and pan-genotype drugs can be used. New treatments are without interferon and do not have the problems of injection, complications, and cost. The cost of new drugs for the patient is relatively low and they do not have many side effects. Currently, hepatitis C can be treated very simply and in most cases by taking only one pill for 12 weeks, and the continuous treatment response (SVR, (Sustained Viral Response) remaining negative in PCR, 12 to 24 weeks after stopping treatment) in more than it happens in 90% of patients. (In people with cirrhosis, despite SVR after definitive treatment, liver ultrasound and FPα measurement are recommended every 6 months to check for liver cancer.

The best strategy to eradicate hepatitis C is to increase the detection and treatment of infected people and stop the cycle of hepatitis C in society. Currently, there is no effective vaccine for hepatitis C. Healthy blood transfusion, compliance with medical matters in hospitals and outpatient clinics, increasing people's awareness about the dangers of tattooing, sexual contact, and injection addiction are among the things that are needed to eradicate this disease.

Decision tree and regression

A decision tree is a tree that categorizes samples in a way that grows from the root downwards and finally reaches the leaf nodes. Each internal or non-leaf node is characterized by a feature, this feature raises a question related to the input example. In each internal node, there are as many possible answers as branches to this question, each of which is determined by the value of that answer. Be the leaves of this tree are defined by a class or a group of answers.

The reason why it is named a decision tree is that this tree shows the decision process for determining the category of an input example. In this research, the category characteristic determines the probability of hepatitis C disease and its types. Figure 2 shows the text view and Figure 3 shows the graphical view of the decision tree applied in this research.

```
PerformanceVector:
accuracy: 94.57%
ConfusionMatrix:
True:   0.0   4.0   1.0   2.0   3.0
0.0:    523   0     13    4     0
4.0:    0     5     0     0     0
1.0:    1     1     7     5     4
2.0:    2     0     0     3     1
3.0:    0     1     0     0     19
classification_error: 5.43%
ConfusionMatrix:
True:   0.0   4.0   1.0   2.0   3.0
0.0:    523   0     13    4     0
4.0:    0     5     0     0     0
1.0:    1     1     7     5     4
2.0:    2     0     0     3     1
3.0:    0     1     0     0     19
kappa: 0.696
ConfusionMatrix:
True:   0.0   4.0   1.0   2.0   3.0
0.0:    523   0     13    4     0
4.0:    0     5     0     0     0
1.0:    1     1     7     5     4
2.0:    2     0     0     3     1
3.0:    0     1     0     0     19
correlation: 0.858
```

Figure 2: Text view of the implemented decision tree.

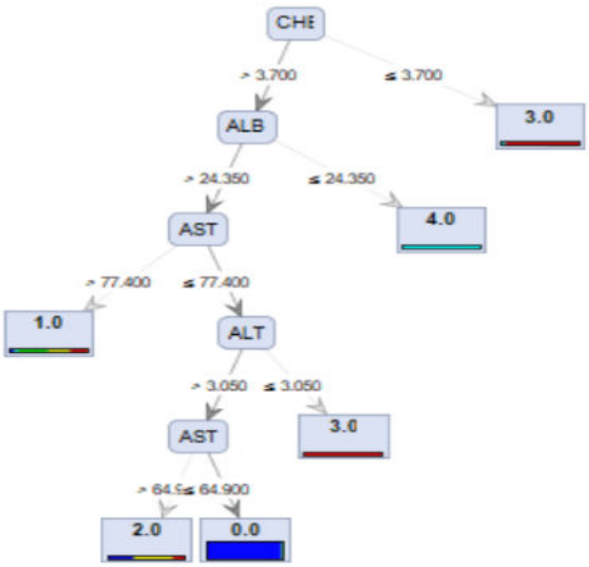

Figure 3: Graphic view of the decision tree of the data set.

The regression technique can be accepted for prediction. Regression analysis can be used to model the relationships of one or more independent and dependent variables. In information extraction, the independent variables are the characteristics that are already known and the dependent variables are related to what we want to predict. Unfortunately, many real-world problems are not easily predicted, so more complex disaggregation may be necessary to predict future values, and a variety of models are often used for regression and classification. In this article, the linear regression modeling method is used for data analysis. Table 1 shows the implemented regression.

Table 1: Output table of total data regression.

Attribute	Coefficient	Std. Error	Std. Coeffici	Tolerance	t-Stat	p-Value	Code
Sex = 0.0	-0.020	0.585	-0.016	0.992	-0.035	0.972	
Sex = 1.0	0.020	0.586	0.026	0.992	0.035	0.972	
Age	0.005	0.002	0.001	0.984	2.207	0.032	**
ALB	-0.029	0.005	-0.004	0.850	-6.170	0	****
ALT	-0.002	0.001	-0.001	0.943	-2,145	0.037	**
AST	0.009	0.001	0.009	0.740	11.375	0	****
BIL	0.010	0.001	0.016	0.872	7.560	0	****
CHE	-0.006	0.012	-0.002	0.825	-0.494	0.626	
CHOL	-0.073	0.021	-0.015	0.922	-3.450	0.001	****
CREA	0.003	0.000	0.002	0.998	6.124	0	****
GGT	0.003	0.000	0.005	0.781	6.852	0	****
PROT	-0.009	0.005	-0.001	0.933	-1.857	0.076	*
(Intercept)	1.696	0.801	?	?	2.116	0.040	**

Note: ALB: Albumin; ALT: alanine aminotransferase; AST: Aminotransferase; BIL: Bilirubin; CHE: Cholinesterase; CHOL: Cholesterol; CREA: Creatinine; PROT: Protein

Test and experiment

The data sets used in this research were selected from the data repository of the Clinical Chemistry Institute of the Medical University of Hannover (MHH). The reason for using these data is that these data were obtained from the laboratory and the statistical population of these data is real patients.

The data is a dataset containing the medical records of 615 blood donor patients collected at the Medical University of Hannover (MHH) and the Helmholtz Center for Infection Research in Braunschweig, Germany. These data include laboratory values of blood donors and hepatitis C patients and demographic values such as age.

This dataset contains 13 features that report clinical body information (Table 2), which are briefly described here:

- A binary attribute related to gender, if it has a value of zero, the patient is male and if it has a value of one, the patient is female.
- An integer feature that specifies the patient's age.
- Category is also a multi-value attribute, if it gets a value of 0, it means that the donor is healthy, if it gets a value of 1, it means that the person has normal hepatitis C, if it gets a value of 2, it means that the person has hepatitis C with fibrosis if it gets a value of 3, it means that the person He has hepatitis C with cirrhosis, and if he gets a value of 4, it means that the donor is suspicious.
- 10 other features are also in decimal form, which includes parameters of ALB, ALP, ALT, AST, BIL, CHE, CHOL, GGT, PROT, and CREA blood.

Table 2: Data characteristics and their values.

Property	Unit of measurement	Description	Period
Age	Age of the patient	Integer	(32…64)
Sex	Gender of the patient	Boolean	(0,1)
ALB	Albumin is a protein made by the liver. Albumin makes up about 60% of the total protein in the blood and plays many roles. This test measures the amount of albumin in the blood	Real	(19.3… 82.2)
ALT	ALT or alanine aminotransferase test is used to measure the level of ALT in the blood. This test is also known as SGPT. ALT is an enzyme made by liver cells	Real	(4….97.8)
AST	Aminotransferase (AST) is an enzyme that exists in various body tissues. An enzyme is actually a type of protein that the body needs to carry out its chemical reactions.	Real	(12….96.2)
BIL	It means checking the amount of bilirubin in the blood. Too much bilirubin in the blood leads to yellowing of the skin and whites of the eyes, which is known as jaundice. Bilirubin is formed in the body when the hemoglobin protein in old red blood cells breaks down.	Real	(2…9.9)
CHE	Serum cholinesterase is a blood test that measures the level of 2 substances that help the nervous system function properly. They are called acetylcholinesterase and pseudocholinesterase. Your nerves need these substances to send signals. Acetylcholinesterase is found in	Real	(2…9.99)

	nerve tissue and red blood cells. Pseudocholinesterase is mainly found in the liver.		
CHOL	Cholesterol is a substance that must be present in the body in a natural and necessary amount. High blood cholesterol leads to the formation of plaque in the arteries; As a result, complications such as arteriosclerosis, hardening of the arteries and finally heart attack follow.	Real	(1.43…9.67)
CREA	Checking the level of creatinine in the blood can easily show the correct or incorrect functioning of the kidneys, because the kidneys are the only ones responsible for removing creatinine from the blood, and if they do not do this properly, it means that they are not functioning properly.	Real	(8…97.7)
GGT	A GGT test is requested to evaluate a possible liver or biliary tract disease or to differentiate between liver and bone disease as the cause of elevated alkaline phosphatase (ALP).	real	(7…99.7)
PROT	It checks the amount of protein in the whole blood. Hemoglobin, enzymes and hormones need this substance to function properly. Protein is obtained through food. Protein will be processed in the liver.	Real	(47….87.7)
(target) Category	It determines the condition of a person with hepatitis C disease and its types	Polynomial	(0….4)

Note: ALB: Albumin; ALT: alanine aminotransferase; AST: Aminotransferase; BIL: Bilirubin; CHE: Cholinesterase; CHOL: Cholesterol; CREA: Creatinine; PROT: Protein

In order to determine the influence of features in the decision tree method, we understand the importance of each feature relative to the root.

In linear regression, by checking the numerical value of the coefficient of each feature in the table, we will understand the importance of that feature [12]. After examining the features repeated in the decision tree as well as the features with high coefficient in the regression tables, we notice the consensus of both methods in the high importance of the three features ALB, AST and CHE. As a result, these three factors are the most important factors of the mentioned diseases and according to the importance of these three characteristics, patients and doctors can prevent the occurrence of hepatitis C and its advanced types by controlling serum albumin, aminotransferase, and cholinesterase in the blood.

Conclusion

In this article, the category of data mining and its techniques were investigated. Also, among the methods of classification and prediction, decision tree and regression were discussed as one of the powerful and common tools in data mining which are easy to understand, implement and use and are computationally cheap. Considering the importance and sensitivity of data mining in medicine, as well as the urgent need of this industry to move from traditional medicine to evidence based medicine, therefore, in this article, the application of data mining in the field of health, especially in identifying the factors in the occurrence of hepatitis C and Advanced types were investigated. In the following, we will examine the results of this implementation:

The following results can be obtained from the analysis of the implemented decision tree (according to Figure 3).

First line: If the CHE characteristic was greater than 3.700, we will examine the ALB characteristic, and if the CHE characteristic was less than or equal to 3.700, the patient has hepatitis C with liver cirrhosis.

Second line: If we are on the left side of the tree from the previous line, we consider the ALB attribute. If the value of this attribute is greater than 24.350, we check the value of the AST attribute. But if the value of this characteristic is less than or equal to this value, this donor is suspicious.

Third line: If we are on the left side of the tree from the previous line, we consider the AST attribute. If the value of this characteristic is more than 77,400, the person is suffering from simple hepatitis C disease. But if the value of this feature is less than or equal to 77,400, we will check the ALT feature.

Fourth line: If we are on the right side of the tree, we consider the ALT attribute. If the value of this attribute is greater than 3.050, we check the value of the AST attribute. But if the value of this feature is less than or equal to 3.050, the patient has hepatitis C with liver cirrhosis.

Fifth line: If we are on the right side of the tree, we consider the AST attribute. If the value of this feature is more than 64,900, the patient has hepatitis C with liver fibrosis. But if the value of this feature is less than or equal to 64,900, the donor is healthy.

The following formula can be obtained from the analysis of the table resulting from the implementation of regression to calculate the characteristics of the category (according to Table 1):

Category= ± 0.020 SEX+0.005 AGE-0.029 ALB-0.002 ALP+0.009 AST+0.010 BIL-0.006 CHE-0.073 CHOL+0.003 CREA +0.003 GGA-0.009 PROT.

Both methods have very high accuracy, but the decision tree with a total accuracy of 94.57% has a higher accuracy than regression with an accuracy of 91.28%. These numbers have been obtained by considering the accuracy of the class in predicting and recalling each feature to bring the algorithm to the positive and negative result of the category feature.

References

1. European Association for the Study of the Liver (EASL) (2018) EASL Recommendations on Treatment of Hepatitis C 2018. J Hepatol 69:461-511
2. Mahmud S, Akbarzadeh V, Abu-Raddad LJ (2018) The epidemiology of hepatitis C virus in Iran: Systematic review and meta-analyses. Sci Rep 8:150
3. Taherkhani R, Farshadpour F (2016) Epidemiology of hepatitis C virus in Iran. World J Gastroenterol 22:5143-53
4. Taherkhani R, Farshadpour F (2017) Global elimination of hepatitis C virus infection: Progresses and the remaining challenges. World J Hepatol 9:1239-1252
5. Alavian SM, Hajarizadeh B, BagheriLankarani K, Sharafi H, Ebrahimi Daryan N, et al. (2016) Recommendations for the Clinical Management of Hepatitis C in Iran: A Consensus-Based National Guideline. Hepat Mon 16:e40959
6. Alavian SM, Sharafi H (2017) Update on Recommendations for the Clinical Management of Hepatitis C in Iran 2017. Hepatmon 17:e63956
7. Bashiri A, Ghazisaeedi M, Saeedfar R, Shahmoradi L, Ehteshami H (2017) Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. Iran J Public Health 46-165-172
8. Buchan TA, Ross HJ, McDonald M, Billia F, Delgado D (2019) Physician prediction versus model predicted prognosis in ambulatory patients with heart failure. J Heart Lung Transplant 38:381
9. Chicco D, Rovelli C (2019) Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PLoS One 14:0208737
10. Chen T, Guestrin C (2016) Xg Boost: a scalable tree boosting system. In: Proceedings of KDD (2016) the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York City: Association for Computing Machinery (ACM). 1-13
11. Rogers G, Joyner E (2011) Mining Your Data for Healthcare Quality Improvement. SAS Institute, Inc., Cary, North Carolina.
12. Balib RK (2005) Clinical Knowledge Management: Opportunities and Challenges. Hershey: Idea Group Inc (IGI), UK. 300