JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 1, January 2026, pp. 15-27

Comparing MARS and Binary Logistic Regression to Modelling Hepatitis C Cases using the SMOTE Balancing Method

Nur Chamidah[1*], Aulia Ramadhanti[1], Azzah Nazhifa Wina Ramadhani[1], Bimo Okta Syahputra[1], Jovansha Ariyawan[1], Ardi Kurniawan[1]
[1]Department of Mathematics, Universitas Airlangga, Indonesia
nur-c@fst.unair.ac.id

Article History:
Received : 15-07-2025
Revised : 29-10-2025
Accepted : 01-11-2025
Online : 01-01-2026

Keywords:
Egypt;
Hepatitis C;
Binary Logistic Regression;
MARS;
SMOTE.

ABSTRACT

Hepatitis is an inflammatory liver disease caused by viral infection and remains a major global public health concern, responsible for approximately 1.4 million deaths annually. Egypt is among the countries with the highest prevalence of Hepatitis C. To address this issue and support Goal 3 of the Sustainable Development Goals (SDGs), this study applies a quantitative approach using secondary data to analyze factors influencing Hepatitis C infection in Egypt. Two statistical models Binary Logistic Regression and Multivariate Adaptive Regression Splines (MARS) were compared, with the SMOTE method implemented to correct class imbalance. The dataset consisted of 608 patient observations, initially imbalanced at a ratio of 86.5:13.5, and were balanced to 52.6:47.4 after SMOTE application. The results revealed that the MARS model demonstrated superior predictive performance compared to binary logistic regression. All independent variables were found statistically significant ($p < 0.05$), except sex. Additionally, all odds ratios were less than 1, indicating a lower probability of Hepatitis C infection relative to non-infection. These findings highlight the relevance of statistical modeling and data-driven strategies in supporting preventive health measures.

https://doi.org/10.31764/jtam.v10i1.33196

A. INTRODUCTION

Hepatitis is an inflammatory liver disease caused by viral infection and is considered contagious. There are five types of viruses that can cause hepatitis: Hepatitis A (HAV), Hepatitis B (HBV), Hepatitis C (HCV), Hepatitis Delta (HDV), and Hepatitis E (HEV) (Green, 2016). Viral hepatitis remains a global public health challenge, responsible for approximately 1.4 million deaths annually worldwide, in which primarily due to complications such as cirrhosis and liver cancer (Devarbhavi et al., 2023). According to the Basic Health Research (Riskesdas) report, noted that 1,017,290 individuals, or around 0.39% of the Indonesian population, were diagnosed with hepatitis in 2018, with the highest prevalence (0.46%) found among individuals aged 45–54 years (Miranda & Adiwinoto, 2022). Beyond viral causes, hepatitis can also result from toxins and chemical exposure, including excessive alcohol consumption and autoimmune diseases (Bataller et al, 2022; Muratori et al., 2022).

Hepatitis C is one of the most prevalent forms of viral hepatitis worldwide. Egypt had the highest prevalence of Hepatitis C globally in 2015, with at least one in five individuals aged 50–59 years infected (The World Bank, 2024). The widespread use of parenteral anti-schistosomal

therapy, which involved frequent reuse of inadequately sterilized syringes, was the primary factor contributing to the high incidence of HCV infection in the past decades (Ayoub et al., 2020). Globally, Hepatitis C remains a critical concern, with more than 290,000 deaths annually and transmission primarily through blood contact, unsafe injections, and non-sterile transfusions (Stroffolini & Stroffolini, 2024). Although direct-acting antiviral (DAA) therapies have proven highly effective, the absence of a vaccine and limited treatment access in developing countries continue to hinder global eradication efforts (Hwang et al., 2025).

Previous studies have examined various methods for analyzing hepatitis. The performance of logistic regression with machine learning algorithms in predicting hepatitis yielded an accuracy rate of 84.62%, suggesting their effectiveness in classifying hepatitis risk (Amrin et al., 2025). In bivariate analysis, education level and wealth index were identified as significant predictors of hepatitis; however, only gender remained significant in the multivariate model (Feliansyah & Purwanto, 2024). However, these studies have not specifically addressed the classification of Hepatitis C status and its associated factors, despite the importance of such mapping for more effective health interventions.

Addressing these research gaps, this study used to identify and model factors associated with Hepatitis C infection. Two statistical techniques, that is Binary Logistic Regression and Multivariate Adaptive Regression Splines (MARS) are utilized to compare their effectiveness in classifying HCV status and determining significant predictors. Unlike previous studies, this research contributes to the literature by integrating a flexible regression framework (MARS) capable of capturing nonlinear relationships among variables and by providing an empirical comparison of model performance for Hepatitis C classification. The findings are expected to enhance understanding of key determinants influencing HCV infection and support the achievement of Sustainable Development Goal (SDG) 3, particularly Target 3.3, which aims to end the epidemics of communicable diseases such as hepatitis by 2030.

B. METHODS

1. Data Source

The type of research used in this study is quantitative research using secondary data, which refers to data obtained from pre-existing studies or sources. The data for this research were collected from the UCI Machine Learning Repository under the dataset titled “Hepatitis C Virus (HCV) for Egyptian patients” consists of medical records of patients diagnosed with hepatitis C at the Faculty of Medicine, Ain Shams University, Egypt, who were undergoing treatment during the year 2017(UCI Machine Learning Repository, 2017).

2. Variables

The dataset includes 607 observations and 13 features or variables, two of which are categorical. Although no previous study has used the Multivariate Adaptive Regression Splines (MARS) method to analyse hepatitis C, previous research has shown that several clinical indicators used in this dataset such as AST, bilirubin, GGT, and cholesterol levels are influenced by gender and are relevant to hepatitis C diagnosis (Dufour et al., 2000; Yi et al., 2019). The variables used in this study are presented in Table 1 below.

Table 1. Research Variables

Variables	Description	Unit	Scale
Y	Hepatitis C disease status	0 = absent 1 = present	Categoric
X_1	Gender	0 = male 1 = female	Categoric
X_2	Aspartate Amino-transferase	U/L	Integer
X_3	Bilirubin	mg/dL	Integer
X_4	Cholesterol	mg/dL	Integer
X_5	Gamma-glutamyl Transferase	U/L	Integer

3. Data Preprocessing

Data preprocessing included applying the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, along with data sampling and splitting into training and testing sets. These steps ensured balanced representation, reduced model bias, and improved generalization and predictive reliability on unseen data (Wongvorachan et al., 2023; Nurhayati & Rahardi, 2024; Haryawan & Ardhana, 2023; Singh et al., 2021).

4. Analysis Method

This study employed Binary Logistic Regression and Multivariate Adaptive Regression Splines (MARS) to model and predict Hepatitis C infection. Binary Logistic Regression was applied to examine relationships between predictor variables and a binary outcome (Anugrawati et al., 2023), while MARS, as a non-parametric method, captured nonlinear interactions with high predictive accuracy (Liu et al., 2023). Both methods were compared to determine the most effective approach for identifying key risk factors.

5. Model Evaluation

To determine the best method, the R^2, AUC, and APPER calculation was used to evaluate classification accuracy. The Apparent Error Rate (APPER) measures the proportion of misclassified data in the training set, providing an initial estimate of model accuracy (Han et al., 2022). The best method was selected based on the lowest APPER value, or the highest accuracy value.

C. RESULT AND DISCUSSION

1. SMOTE Balancing

This study uses Hepatitis C patient data classified into two categories, namely uninfected (code 0) and infected (code 1), based on the diagnosis results as the response variable with the number of each class as follows in Figure 2.

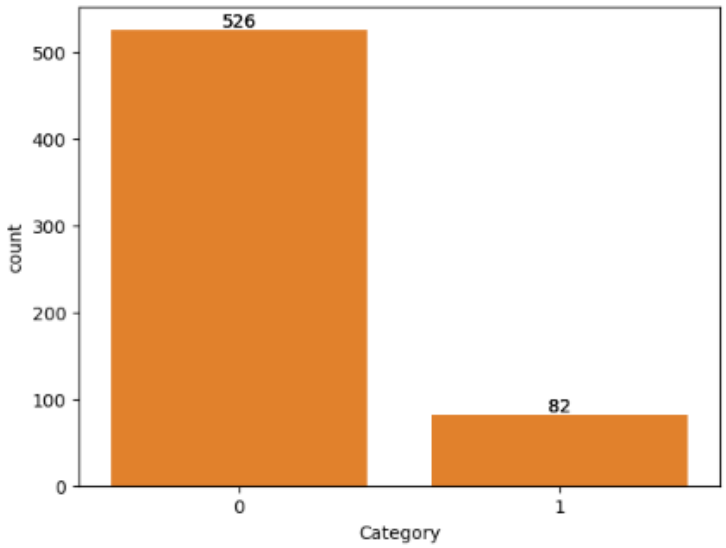

Figure 1. Number of Hepatitis C Patients for Each Code in the Initial Data

Figure 1 shows the imbalance in the number of patients in each category, with 526 uninfected patients and 82 infected patients. This imbalance reflects the uneven distribution of the baseline data. To overcome the data imbalance, balancing is done using the SMOTE technique with machine learning so that the amount of data in both categories becomes balanced as follows in Figure 2.

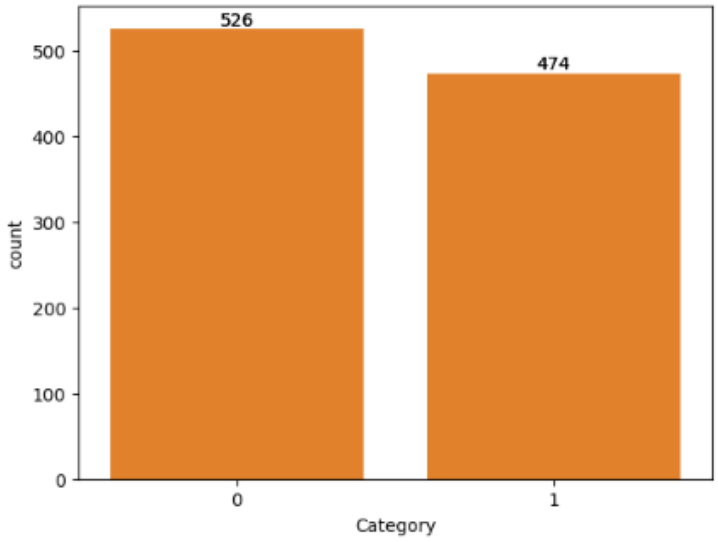

Figure 2. Number of Synthesised Data for Each Code

The total amount of data in this study is 1,000 observations, which are divided into training data with 900 observations and testing data with 100 observations to train and test the model objectively.

2. Descriptive Statistics

After obtaining sample data, the characteristics for each variable for training data can be found as follows Table 2.

Table 2. Characteristics of Variables

Variable	N	Mean	StDev	Min	Max
X_2	900	55.728	50.816	10.6	324
X_3	900	17.383	29.559	0.8	254
X_4	900	5.076	1.144	1.43	9.67
X_5	900	72.631	99.676	4.5	650.9

Based on Table 2 presents a description of the research data for each variable, the number of samples (N) totalling 900, the mean value (Mean), standard deviation (StDev), minimum value (Min), and maximum value (Max) are presented. For example, variable X_2 has a mean of 55.728 with a data spread measured by a standard deviation of 50.816, and its data values range from 10.6 to 324. Similarly, this statistical description provides a concise overview of the data characteristics for variables X_3, X_4, and X_5 in this study.

3. Binary Logistic Regression Analysis

As a first step, multicollinearity detection was performed on the simulated data between each predictor variable. The results of multicollinearity testing based on VIF values in Table 3.

Table 3. VIF Value for Predictor Variables

	X_1	X_2	X_3	X_4	X_5
VIF	1.111	1.505	1.099	1.142	1.330

Based on Table 3, the multicollinearity test results show that all VIF values for variables X_1 to X_5 are below 10, which is between 1.099 and 1.505. This indicates that there is no significant multicollinearity problem, so the multicollinearity assumption in the model has been met.

a. Model Fit Test

The binary logistic model fit test is used to compare the actual model with the predicted model. Based on the results of the Hosmer and Lemeshow Test, the p-value is 0.874 where the result is > α (0.05), then the model decision is obtained, meaning that the binary logistic regression model is suitable for modelling data on the effect of Sex, AST, BIL, CHOL, and GGT on the diagnosis of Hepatitis C disease because there is no significant difference between the predicted probabilities and the observed probabilities.

b. Parameter Overall Test

To determine whether any independent variables influence the model, the overall parameter test results are presented in Table 4.

Table 4. Result of the Overall Test

Model	Chi-Square	df	p-Value	Decision
Final	807.048	5	0.000	H_0 rejected

From Table 4, the p-value is 0.000 where the result is < α (0.05), then H_0 is rejected, which means that at least one of the variables Sex, AST, BIL, CHOL, and GGT has an effect on the diagnosis of Hepatitis C disease.

c. Parameter Partial Test

Partial parameter testing is conducted to assess the significance of each independent variable in the binary logistic regression model. The results are presented in Table 5.

Table 5. First Stage Wald Test

Variable	Wald	p-Value	Decision
X_1	0.010	0.921	H_0 accepted
X_2	73.633	0.000	H_0 rejected
X_3	10.988	0.001	H_0 rejected
X_4	26.416	0.000	H_0 rejected
X_5	37.199	0.000	H_0 rejected

From Table 5, the results of the partial effect test of each predictor variable on the response variable are as follows.

1) The significance value of variable X_1 or the Sex variable is 0.921 where the result is > α, then H_0 is accepted, which means that gender has no effect on the diagnosis of Hepatitis C.
2) The significance value of variable X_2 or the AST (Aspartate Amino-transferase) variable is 0.000 where the result is < α, then H_0 is rejected, which means that the amount of AST (Aspartate Amino-transferase) affects the diagnosis of Hepatitis C disease.
3) The significance value of variable X_3 or the BIL (Bilirubin) variable is 0.001 where the result is < α, then H_0 is rejected, which means that the amount of BIL (Bilirubin) affects the diagnosis of Hepatitis C disease.
4) The significance value of variable X_4 or the CHOL (Cholesterol) variable is 0.000 where the result is < α, then H_0 is rejected, which means that the CHOL (Cholesterol) level affects the diagnosis of Hepatitis C.
5) The significance value of variable X_5 or the GGT (Gamma-glutamyl Transferase) variable is 0.000 where the result is < α, then H_0 is rejected, which means that the amount of GGT (Gamma-glutamyl Transferase) affects the diagnosis of Hepatitis C disease.

Based on the partial test results, it is found that the variables AST, BIL, CHOL GGT have a partially significant effect on the diagnosis of Hepatitis C disease. Therefore, re-modeling will be carried out to obtain the best model by eliminating the predictor variable that does not significantly affect the diagnosis of Hepatitis C disease, namely the Sex variable. By considering the predictor variables that influence the diagnosis of Hepatitis C disease, the results of simultaneous and partial testing are obtained as follows Table 6.

Table 6. Second Stage Wald Test

Variable	B	Wald	p-Value	Decision
X_2	0.107	75.220	0.000	H_0 rejected
X_3	0.060	11.184	0.001	H_0 rejected
X_4	-0.659	26.680	0.000	H_0 rejected
X_5	0.023	37.500	0.000	H_0 rejected
Constant	-2.806	12.661	0.000	H_0 rejected

d. Binary Logistic Regression Modelling and Odds Ratio

Based on the results in Table 6, the binary logitistic model formed is as follows.

$$\ln\left(\frac{\exp(\boldsymbol{X_i\beta})}{1+\exp(\boldsymbol{X_i\beta})}\right) = \alpha + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$
$$= -2.806 + 0.107X_2 + 0.060X_3 - 0.659X_4 + 0.023X_5$$

While the odds ratio is used to facilitate the interpretation of the binary logistic regression model, where it is known from the results of the partial parameter significance test that the variables AST, BIL, CHOL, and GGT each significantly affect the diagnosis of Hepatitis C. The odds ratio value of each variable that has a significant effect on the diagnosis of Hepatitis C can be seen in Table 7 below.

Table 7. Odd Ratio Value for Each Variable

Variable	Exp($\boldsymbol{\beta}$)
X_2	1.113
X_3	1.062
X_4	0.518
X_5	1.023

1) $OR_2 = \exp(\beta_2) = 1.113$

This means that the odd ratio value or exp(β_2) of the AST (Aspartate Amino-transferase) of 1.113. This indicates that if there is an increase in AST (Aspartate Amino-transferase) by 1 unit, it is estimated that a person is 1.113 times more likely to be diagnosed with Hepatitis C disease.

2) $OR_3 = \exp(\beta_3) = 1.062$

This means that the odd ratio or exp(β_3) value of the BIL (Bilirubin) variable is 1.062. This indicates that if there is an increase in BIL (Bilirubin) by 1 unit, it is estimated that a person is 1.062 times more likely to be diagnosed with Hepatitis C disease.

3) $OR_4 = \exp(\beta_4) = 0.518$

This means that the odd ratio or exp(β_4) of the CHOL variable is equal to 0.518. This indicates that if there is an increase in CHOL (Cholesterol) levels by 1 unit, it is estimated that a person is 0.518 times more likely to be diagnosed with Hepatitis C disease.

4) $OR_5 = \exp(\beta_5) = 1.023$

This means that the odd ratio or exp(β_5) of the variable GGT (Gamma- glutamyl Transferase) of 1.023. This indicates that if there is an increase in GGT (Gamma-

glutamyl Transferase) by 1 unit, it is estimated that a person is 1.023 times more likely to be diagnosed with Hepatitis C disease.

e. Probability Estimation and Coefficient Determination
The binary logistic regression equation to determine the estimated probability of being diagnosed with Hepatitis C disease is as follows.

$$\pi_i = P(Y=1) = \frac{\exp(\alpha + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}{1 + \exp(\alpha + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}$$

$$P(Y=1) = \frac{\exp(-2.806 + 0.107X_2 + 0.060X_3 - 0.659X_4 + 0.023X_5)}{1 + \exp(-2.806 + 0.107X_2 + 0.060X_3 - 0.659X_4 + 0.023X_5)}$$

with R Square result is 0.790, which shows that the ability of the predictor variables, namely AST, BIL, CHOL, and GGT, to explain the diagnosis of Hepatitis C disease, is 79% and the rest is explained by other variables not included in the model.

f. Classification Accuracy
Evaluating classification accuracy in the accuracy of classification of categories non-hepatitis and affected by hepatitis C can use the Apparent Error Rate (APER). Classification accuracy output based on categories as shown in Table 8.

Table 8. Classification Accuracy Result

Observed		Predicted		
		Category		Percentage Correct
		Non-Hepatitis	Hepatitis	
Category	Non-Hepatitis	445	27	94.3
	Hepatitis	58	370	86.4
Overall Percentage				90.6

Based on Table 8, it is known that the number of patients who is non-hepatitis is 445 patients, 27 of whom are predicted to affect from Hepatitis C with a prediction rate of 94.3%. While it is known that the number of patients affecting from Hepatitis C is 370 patients, 58 patients among them are predicted not to affect from Hepatitis C with a prediction rate of 86.4%. Thus, the percentage of classification accuracy of the model can predict correctly by 90.6%. If calculated manually, the APER values are as follows.

$$\text{APER} = \frac{27+58}{445+27+58+370} = 0.094$$

with classification accuracy or accuracy can be formulated as 1– APER = 1– 0.094 = 0.906, so it is true that the similarity between manual calculations and output from using SPSS *software* is 90.6%.

g. Prediction with the best Binary Logistic Regression model
Based on the binary logistic regression model that was formed, predictions were then made on the testing data using the modeling results on the training data from the best

binary logistic regression model. The following is an evaluation matrix for the test data, as shown in Table 9.

Table 9. Classification Accuracy of Binary Logistic Regression Predictions

Observed		Predicted Category	
		Non- Hepatitis	**Hepatitis**
Category	Non-Hepatitis	51	3
	Hepatitis	3	43

Based on the output in Table 9, the results of the Sensitivity calculation are 0.9348, Specificity is 0.9444, Negative Predictive Value is 0.9444, Positive Predictive Value is 0.9348, and the APER calculation is as follows.

$$\text{APER} = \frac{3+3}{43+3+3+51} = 0.06$$

With classification accuracy or accuracy can be formulated as 1− APER = 1− 0.06 = 0.94, so we get a classification accuracy of 94% for the *testing* data. In addition, an AUC value of 0.9396 was obtained, indicating that the binary logistic regression model has a good ability to distinguish between positive and negative classes.

4. MARS Analysis

Based on the trial-and-error method to examine the best basis function combination to get the best MARS model. It was obtained that BF = 20, MI = 2, and MO = 1 as the best combination, since it had the most maximum R^2 value, that is 0.771. The next step is to determine the importance level of each predictor variables, which may affect the model form. The test using MARS software given in Table 10 below.

Table 10. Importance Level of Each Predictor Variable

Variable	Importance	-GCV
X_2	100.000	0.109
X_3	48.787	0.074
X_5	46.068	0.073
X_4	32.168	0.068
X_1	0.000	0.063

Based on the Table 10, it is known that X_1 has the importance of 0.000 which means that Sex variable has no influence in the model obtained. Meanwhile, the other variable has their importance throughout the model, even though in different levels. After knowing the importance of each variable, it is needed to know the influence of each basis functions to the model. The result is given in Table 11 below.

Table 11. Simultaneous Test of MARS Basis Function

F	$F_{(0.05;14,885)}$	p-Value	Decision
212.311	0.468	0.000	H_0 rejected

Table 11 shows that the F value is 212.311 with the p-value 0.000. Since the F value obtained is greater than F table and the p-value is under the critical area of 0.05, it can be concluded that there was a significant influence simultaneously of all the basic functions in MARS model obtained to the response variable used. Then, the next step is to examine the MARS model based on the coefficient of the basic model obtained, which given in Table 12 below.

Table 12. Coefficient of MARS Model Basis Functions

Parameter	Estimate	S.E.	T-Ratio	p-Value
Constant	2.371	0.093	25.617	0.000
BF 1	0.060	0.006	10.130	0.000
BF 2	-0.054	0.003	-15.813	0.000
BF 4	-0.006	0.001	-8.589	0.000
BF 5	-0.111	0.014	-7.710	0.000
BF 7	-0.072	0.005	-13.412	0.000
BF 8	0.001	0.000	8.755	0.000
BF 10	-0.062	0.006	10.577	0.000
BF 12	0.000	0.000	5.400	0.000
BF 13	0.000	0.000	4.848	0.000
BF 14	-0.005	0.001	-5.119	0.000
BF 16	0.005	0.001	3.937	0.000
BF 17	0.004	0.000	10.077	0.000
BF 18	-0.003	0.000	-4.075	0.000
BF 20	0.001	0.000	4.055	0.000

It can be shown from Table 12, above that all basis function has a significant influence to the model obtained, since the p-value is under 0.05. The equation of each basis function may be given below.

Table 13. Basis Function Equation for the Selected MARS Model

Basis Function	Equation
BF 1	$= \max(0; X_2 - 54)$
BF 2	$= \max(0; 54 - X_2)$
BF 4	$= \max(0; 131.31 - X_5)$
BF 5	$= \max(0; X_4 - 1.43)$
BF 7	$= \max(0; 13.719 - X_3)$
BF 8	$= \max(0; X_2 - 52.6) * \max(0; 13.719 - X_3)$
BF 10	$= \max(0; X_2 - 41.041)$
BF 12	$= \max(0; 41.041 - X_2)$
BF 13	$= \max(0; X_2 - 27.7) * \max(0; 131.31 - X_5)$
BF 14	$= \max(0; 27.7 - X_2) * \max(0; 131.31 - X_5)$
BF 16	$= \max(0; X_2 - 142.578) * \max(0; 13.719 - X_3)$
BF 17	$= \max(0; X_3 - 14.8) * \max(0; 41.041 - X_2)$
BF 18	$= \max(0; 14.8 - X_3) * \max(0; 41.041 - X_2)$
BF 20	$= \max(0; X_2 - 14.429) * \max(0; 54 - X_2)$

Based on the Table 13 above, it can be written the best MARS model which given as follows.

$$Y = 2.371 + 0.060 * BF1 - 0.054 * BF2 - 0.006 * BF4 - 0.111 * BF5 - 0.072 * BF7 + 0.001 * BF8 - 0.062 * BF10 + 0.00003 * BF12 + 0.00002 * BF13 - 0.005 * BF14 + 0.005 * BF16 + 0.004 * BF17 - 0.003 * BF18 + 0.001 * BF20$$
$$= 2.371 + 0.060 * \max(0; X_2 - 54) - 0.054 * \max(0; 54 - X_2) - 0.006 * \max(0; 131.31 - X_5) - 0.111 * \max(0; X_4 - 1.43) - 0.072 * \max(0; 13.719 - X_3) + 0.001 * \max(0; X_2 - 52.6) * \max(0; 13.719 - X_3) - 0.062 * \max(0; X_2 - 41.041) + 0.00003 * \max(0; X_2 - 27.7) * \max(0; 131.31 - X_5) + 0.00002 * \max(0; 27.7 - X_2) * \max(0; 131.31 - X_5) - 0.005 * \max(0; X_2 - 142.578) * \max(0; 13.719 - X_3) + 0.005 * \max(0; X_3 - 14.8) * \max(0; 41.041 - X_2) + 0.004 * \max(0; 14.8 - X_3) * \max(0; 41.041 - X_2) - 0.003 * \max(0; X_2 - 14.429) * \max(0; 54 - X_2) + 0.001 * \max(0; X_4 - 1.43) * \max(0; 131.31 - X_5)$$

Just like the analysis using binary logistic regression above, the MARS model obtained needed to be evaluated whether it is good to use as the classification method, as shown in Table 14.

Table 14. Classification Accuracy of the MARS Method

Observed		Predicted		
		Category		Percentage Correct
		Non-Hepatitis	Hepatitis	
Category	Non-Hepatitis	454	18	96.19%
	Hepatitis	46	382	89.25%
Overall Percentage				92.89%

Based on Table 14, it is known that there are 454 non-hepatitis patients, which 18 of whom are predicted to have Hepatitis C with a prediction rate of 96.19%. Besides, it is known that the number of patients infected by Hepatitis C is 382 patients, with 46 patients among them are predicted as a non-Hepatitis patients with a prediction rate of 89.25%. Thus, the percentage of classification accuracy of the model can predict correctly by 92.89%.

5. Selecting The Best Regression Method

Based on the analysis result that has been explained above, it can be given a table of evaluation metrics comparison between binary logistic regression and MARS method to examine the best method to modeling the research data, as shown in Table 15.

Table 15. Comparison of Each Regression Method

Method	R^2	Accuracy	AUC
Binary Logistic Regression	0.790	0.906	0.908
MARS	0.771	0.929	0.932

Table 15 shows that MARS have a better prediction performance based on the accuracy and AUC value, even the R^2 value is a little bit smaller than binary logistic regression. This means that MARS model is more effective to predict non-hepatitis and hepatitis patient accurately, and have a lower classification error.

D. CONCLUSION AND SUGGESTIONS

Based on the analysis that has been conducted above, it can be concluded that there was an unbalance amount of the real research data used, so it is needed to be balanced using SMOTE method. The synthesis data after rebalancing is 526 data for non-hepatitic patients and 474 data for hepatitic patients. Another result based on the analysis is that Multivariate Adaptive Regression Spline (MARS) method is better in modeling the data, since it gave an accuracy of 92.9% for the data classification, which categorized as highly accurate. Furthermore, all odds ratios from the MARS model's basis functions were below 1, indicating that the possibility of Hepatitis C infection was generally lower than of not being infected. Future research should consider incorporating additional variables that may influence Hepatitis C incidence to enhance the model's predictive capability and support more targeted prevention strategies. From a practical standpoint, promoting healthy lifestyle practices remains essential to reduce the risk of Hepatitis C infection at the population level.

ACKNOWLEDGEMENT

We extend our gratitude to the Department of Mathematics, Universitas Airlangga, for the resources and support provided throughout this research. We also appreciate the contributions of various authors and researchers whose work has significantly informed the theoretical and methodological foundations of this study. Lastly, we thank everyone who contributed data and offered assistance, enabling the successful completion of this research.

REFERENCES

Amrin, A., Rudianto, R., & Sismadi, S. (2025). Data Mining with Logistic Regression and Support Vector Machine for Hepatitis Disease Diagnosis. *JITE (Journal of Informatics and Telecommunication Engineering), 8*(2), 248–256. https://doi.org/ 10.31289/jite.v8i2.13218

Anugrawati, S. D., Nurhikma, Iyut Wahyu Saputri, & Khalilah Nurfadilah. (2023). Analisis Regresi Logistik Biner dalam Penentuan Faktor-Faktor yang Mempengaruhi Ketepatan Waktu Lulus Mahasiswa UIN Alauddin Makassar. *Journal of Mathematics: Theory and Applications*, *5*(1), 11–16. https://doi.org/10.31605/jomta.v5i1.2401

Ayoub, H. H., Chemaitelly, H., Kouyoumjian, S. P., & Abu-Raddad, L. J. (2020). Characterizing the historical role of parenteral antischistosomal therapy in hepatitis C virus transmission in Egypt. *International Journal of Epidemiology*, *49*(3), 798–809. https://doi.org/10.1093/ije/dyaa052

Bataller, R., Arab, J. P., & Shah, V. (2022). Alcohol-Associated Hepatitis. *The New England Journal of Medicine*, *387*(26), 2436–2448. https://doi.org/10.1056/NEJMra2207599

Devarbhavi, H., Asrani, S. K., Arab, J. P., Nartey, Y. A., Pose, E., & Kamath, P. S. (2023). Global burden of liver disease: 2023 update. *Journal of Hepatology*, *79*(2), 516–537. https://doi.org/10.1016/j.jhep.2023.03.017

Dufour, D. R., Lott, J. A., Nolte, F. S., Gretch, D. R., Koff, R. S., & Seeff, L. B. (2000). Diagnosis and Monitoring of Hepatic Injury. I. Performance Characteristics of Laboratory Tests. *Clinical Chemistry*, *46*(12), 2027–2049. https://doi.org/10.1093/clinchem/46.12.2027

Feliansyah, A. W., & Purwanto, E. (2024). Analisis faktor yang berhubungan dengan penyakit hepatitis di Indonesia. *Holistik Jurnal Kesehatan*, *18*(9), 1131–1138. https://doi.org/10.33024/hjk.v18i9.587

Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques* (4th Ed.). Elsevier.

Haryawan, C., & Ardhana, Y. M. K. (2023). Analisa Perbandingan Teknik Oversampling SMOTE pada Imbalanced Data. *Jurnal Informatika dan Rekayasa Elektronik*, *6*(1), 73–78. https://doi.org/10.36595/jire.v6i1.834

Heneghan, M. A., & Lohse, A. W. (2025). Update in clinical science: Autoimmune hepatitis. *Journal of Hepatology, 82*(5), 926–937. https://doi.org/10.1016/j.jhep.2024.12.041

Hwang, S. Y., Danpanichkul, P., Agopian, V., Mehta, N., Parikh, N. D., Abou-Alfa, G. K., Singal, A. G., & Yang, J. D. (2025). Hepatocellular carcinoma: updates on epidemiology, surveillance, diagnosis and treatment. *Clinical and Molecular Hepatology*, *31*(Suppl), S228–S254. https://doi.org/10.3350/cmh.2024.0824

Liu, Y., Li, D., & Xia, Y. Dimension Reduction and MARS. *JMLR*, *24*(309),1–30. http://jmlr.org/papers/v24/22-1422.html

Miranda, S., & Adiwinoto, R. P. (2022). Tinjauan Sistematik: Epidemiologi Hepatitis A pada Anak di Indonesia. *Prominentia Medical Journal*, *3*(2), 40–55. https://doi.org/10.37715/pmj.v3i2.3216

Nurhayati, L. D., & Rahardi, M. (2025). Impact of SMOTE and ADASYN on Class Imbalance in Metabolic Syndrome Classification Using Random Forest Algorithm. Journal of Applied Informatics and Computing, 9(5), 2807–2813. https://doi.org/10.30871/jaic.v9i5.10657

Singh, V., Pencina, M., Einstein, A. J., Liang, J. X., Berman, D. S., & Slomka, P. (2021). Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific reports*, *11*(1), 14490. https://doi.org/10.1038/s41598-021-93651-5

Stroffolini, T., & Stroffolini, G. (2024). Prevalence and Modes of Transmission of Hepatitis C Virus Infection: A Historical Worldwide Review. *Viruses*, *16*(7), 1115. https://doi.org/10.3390/v16071115

The World Bank. (2024). *How Egypt Won its Battle Against Hepatitis C*. https://www.worldbank.org/en/news/feature/2024/04/05/how-egypt-won-its-battle-against-hepatitis-c

UCI Machine Learning Repository. (2017). *Hepatitis C Virus (HCV) for Egyptian patients - UCI Machine Learning Repository*. https://archive.ics.uci.edu/dataset/503/hepatitis+c+virus+hcv+for+egyptian+patients

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Information, *14*(1), 54. https://doi.org/10.3390/info14010054

Yi, S.-W., Yi, J.-J., & Ohrr, H. (2019). Total cholesterol and all-cause mortality by sex and age: a prospective cohort study among 12.8 million adults. *Scientific Reports*, *9*(1), 1596. https://doi.org/10.1038/s41598-018-38461-y