JAMBURA JOURNAL OF MATHEMATICS
Jambura J. Math. Vol. 5, No. 2, pp. 289–304, August 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom
DOI: https://doi.org/10.34312/jjom.v5i2.20289

The Comparison between Ordinal Logistic Regression and Random Forest Ordinal in Identifying the Factors Causing Diabetes Mellitus

Assyifa Lala Pratiwi Hamid[1,*], Anwar Fitrianto[1], Indahwati[1], Erfiani[1], Khusnia Nurul Khikmah[1]

[1]*Department of Statistics, Faculty of Mathematic and Natural Science, IPB University, Bogor, Indonesia*
[*]*Corresponding author. Email: assyifa.hamid@gmail.com*

ABSTRACT

Diabetes is one of the high-risk diseases. The most prominent symptom of this disease is high blood sugar levels. People with diabetes in Indonesia can reach 30 million people. Therefore, this problem needs further research regarding the factors that cause it. Further analysis can be done using ordinal logistic regression and random forest. Both methods were chosen to compare the modelling results in determining the factors causing diabetes conducted in the CDC dataset. The best model obtained in this study is ordinal logistic regression because it generates an accuracy value of 84.52%, which is higher than the ordinal random forest. The four most important variables causing diabetes are body mass index, hypertension, age, and cholesterol.

Keywords:
Ordinal Logistic Regression; Ordinal Random Forest; Diabetes

Citation Style:

A. L. P. Hamid, *et al.*, "The Comparison between Ordinal Logistic Regression and Random Forest Ordinal in Identifying the Factors Causing Diabetes Mellitus", *Jambura J. Math.*, vol. 5, No. 2, pp. 289–304, 2023, doi: https://doi.org/10.34312/jjom.v5i2.20289

1. Introduction

Diabetes Mellitus is a disease caused by the immune system (immunity) of the patient's body, which ruins and attacks pancreatic cells that work to produce insulin. This can cause an increase in blood glucose levels, which causes damage to the internal organs of the body. Diabetes mellitus includes non-communicable diseases, but the mortality level caused by this disease is one of the highest in the world [1]. Diabetes not only causes premature death around the world, but also becomes the main cause of blindness, cardiac disease, and kidney failure [2]. International Diabetes Federation (IDF) estimated that at least 463 million people in 20-79 years old in the world have diabetes in 2019 or equivalent to the prevalence of 9.3% of the total population at the same age [3].

One of the attempts to decrease the prevalence of Diabetes Mellitus is by understanding the factors causing Diabetes Mellitus, which then we can avoid it. According to the Infodatin 2020 published by the Ministry of Health of the Republic of Indonesia, the diagnosis of Diabetes Mellitus is performed by measuring the blood sugar level, then

e-ISSN: 2656-1344
Received: 2023-06-07 | Revised: 2023-06-27 | Accepted: 2023-07-01| Published: 2023-08-01

the results of the health check will classify whether this person is normal, pre-diabetes, or diabetes. The Infodatin 2020 [4] also stated that the risk factors of diabetes consist of modifiable factors and non-modifiable factors. Non-modifiable factors are race, ethnicity, age, gender, family history of diabetes mellitus, history of giving birth to a baby >4.000 grams, and history of low birth weight (newborn or < 2.500 grams). Modifiable factors are overweight, abdominal/central obesity, lack of physical activity, lack of consuming fruits and vegetables, hypertension, dyslipidemia, unhealthy and unbalanced diet (high calories), and smoking. According to the study conducted by Cahyono and Purwanti [5], another factor is alcohol consumption, in which they also studied the educational level of diabetics. Anggraini [6] stated that cholesterol is also often correlated with diabetes. Moreover, Putra [7] conducted a study regarding how hypertension in diabetics.

The previous study regarding the comparison of random forest ordinal and ordinal logistic regression has been conducted by Nisa, et al. [8] with the topic of identifying factors influencing the achievement of IPB Students in 2022. The other study regarding the comparison of random forest and logistic regression has been conducted by Tanujaya [9] and [10] where in this study random forest has higher accuracy to logistic regression.

Based on the explanation above, research related to the disease and the factors that cause diabetes mellitus needs to be done. This study also proposes a comparison of the performance of the latest methods between ordinal logistic regression and ordinal random forest in order to get the best prediction of the data and find out the performance of the best method in predicting the data. The data used in this study comes from The Behavioral Risk Factor Surveillance System (BRFSS), which is a health-related telephone survey that is collected annually by the Centers for Disease Control (CDC), where this paper structurally helps four sections there. namely introduction, method, results and discussion, and finally the conclusion.

2. Methods

2.1. Research Stages

This study used secondary data retrieved from The Behavioral Risk Factor Surveillance System (BRFSS), which was health-related telephone surveys collected every year by the Centers for Disease Control (CDC) and last updated in 2021. The data is open and publicly accessible for research and scientific development through the following link www.kaggle.com. Data consisting of response variables symbolized by Y and independent variables symbolized by X. Response variables used were the data of observed diabetes status. Independent variables used were hypertension (X1), cholesterol (X2), BMI (X3), smoking (X4), physical activities (X5), fruit consumption (X6), vegetable consumption X7), alcohol consumption (X8), gender (X9), age (X10), and education level (X11). Furthermore, the variables used in this study are shown in Table 1.

Data analysis in this study used R software with the package, MASS, caret, rpart, party, and ordinalForest. The stages in data analysis conducted in this study were as follows:

1. Collected and inputted data.
2. Conducted descriptive statistical analysis of the data.
3. Built ordinal logistic regression model, the stages were:

Table 1. The variables used in the research

Variable	Caterory	Dummy	Reference
Diabetes	No Diabetes	0	[4]
	Prediabetes	1	
	Diabetes	2	
High Blood Pressure	No High Blood Pressure	0	[7]
	High Blood Pressure	1	
High Cholesterol	No High Cholesterol	0	[6]
	High Cholesterol	1	
BMI	Body Mass Index	-	[4]
Smoker	Never smoked 100 cigarettes in a lifetime	0	[4]
	Smoked $\geq$100 cigarettes in a lifetime	1	
Physical Activity	Not doing physical activity not including job in past 30 days	0	[4]
	Doing physical activity not including job in past 30 days	1	
Fruit	Not consume fruit 1 or more times per day	0	[4]
	Consume fruit 1 or more times per day	1	
Veggies	Not consume vegetables 1 or more times per day	0	[4]
	Consume vegetables 1 or more times per day	1	
Heavy Alcohol Consump	Not having more than 14 drinks per week	0	[5]
	Having more than 14 drinks per week	1	
Sex	Female	0	[4][7]
	Male	1	
Age	18-24	1	[4]
	25-29	2	
	30-34	3	
	35-39	4	
	40-44	5	
	45-49	6	
	50-54	7	
	55-59	8	
	60-64	9	
	65-69	10	
	70-74	11	
	75-79	12	
	$\geq$80	13	
Education	Never attended school/ Kindergarden	1	[4]
	Elementary	2	
	Junior High School	3	
	Senior High School	4	
	Graduate	5	
	Post Graduate	6	

(a) Estimated ordinal logistic regression parameters.
(b) Conducted parameter testing simultaneously to find out the roles of all explanatory variables in the model.
(c) Conducted parameter testing partially to find out the explanatory variables

that had a significant impact on the model.
   (d) Conducted model fit test.
   (e) Looked for odds ratio to interpret ordinal logistic regression.
4. Looked for classification accuracy of the ordinal logistic regression.
5. Conducted cross-validation in ordinal logistic regression using 80% training data and 20% test data repeated 100 times.
6. Formed random forest ordinal with 500 trees and three sorting variables.
7. Conducted cross-validation in random forest ordinal using 80% training data and 20% test data repeated 100 times. Training data was also used to determine the importance level of explanatory variables.
8. Compared the evaluation results of ordinal logistic regression and random forest ordinal.

2.1.1. Ordinal Logistic Regression

The statistical method that can describe the correlation between independent variables and response variables, where response variables are more than two categories and its measurement scale is in the form of level, is ordinal logistic regression [11]. The logit model is a model used for ordinal logistic regression. The logit model is the implementation of the GLMs model whose connection function is cumulative logit models. The distribution equation in the category of the response variable with ordinal scale was m, where $r = 1 < ... < m$ that had a characteristic of $\pi_1 + \cdots + \pi_m = 1$. The cumulative distribution function of the response variables of Y was $P\left(Y \leq r \mid X_j\right) = \pi_j(x)$ [12]. It was defined as follows:

$$\pi_j\left(x\right) = P\left(Y \leq r \mid X_j\right) = \frac{\exp(\beta_{0r} + \sum_{j=1}^{P} \beta_j x_{ij})}{1 + \exp(\beta_{0r} + \sum_{j=1}^{P} \beta_j x_{ij})} \tag{1}$$

Furthermore, equation (1) was transformed to a linear function using the logit link function of $Logit\left[\pi,(x)\right] = ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$ as follows:

$$g\left(x\right) = Logit\left(\pi,(x)\right) = \beta_{0r} + \sum_{j=1}^{P} \beta_j x_{ij} + \varepsilon_{ij} \tag{2}$$

Using Maximum Likelihood Estimation method, an estimated value of β_j was obtained that maximizes the function of $l\left(\beta\right) = \prod_{i=1}^{n}\left[\pi_0(x)^{y_{0i}}\pi_1(x)^{y_{1i}}\ldots\pi_m(x)^{y_{mi}}\right]$. The distribution of each response variable was calculated according to the difference in cumulative distribution value of each logit function, which was $\pi_1\left(x\right) = P\left(Y \leq 1\right)$, $\pi_2\left(x\right) = P\left(Y \leq 2\right) - \left(Y \leq 1\right)$ until $P\left(Y \leq m\right) = 1 - P\left(Y \leq m-1\right)$.

2.1.2. Parameter Testing

Parameter testing simultaneously aims to find the influence of the variable jointly on the response variable using Likelihood Ratio Test (G^2). The hypothesis was $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs $H_1 : \beta_r \neq 0,\ r = 1, 2, \ldots, p$ [13]. The statistical test was as follows:

$$G^2 = -2\ln\left[\frac{\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1}\ldots\left(\frac{n_m}{n}\right)^{n_m}}{\prod_{i=1}^{n}\pi_0(x)^{y_{0i}}\pi_1(x)^{y_{1i}}\ldots\pi_m(x)^{y_{mi}}}\right] \tag{3}$$

The decision to reject H_0 was $G^2 > \chi^2_{(\alpha,df)}$ or $P - value < \alpha$ where df is the number of parameters.

Parameter testing partially aims to find the influence of each independent variable on the response variable using the Wald test [14]. The hypothesis was $H_0 : \beta_r = 0,\ r = 1, 2, \ldots, p$ vs $H_1 : \beta_r \neq 0,\ r = 1, 2, \ldots, p$. The statistical test was as follows:

$$W = \frac{\widehat{\beta}_r}{SE(\widehat{\beta})} \tag{4}$$

Decision to reject H_0 if $|W| = Z_{\alpha/2} >$ or $P - value < \alpha$.

2.1.3. Model Fit Test

Data modeled using ordinal logistic regression was then measured by the fit of the data with the model. The fit test used in this analysis was the Hosmer Lemeshow test [15].

$$\chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - \frac{E_g}{n_g})} \tag{5}$$

where O_g signifies the observed events, E_g signifies the expected events and ng signifies the number of observations for the gth group, and G is the number of groups. The test statistic follows a Chi-squared distribution with G^2 degrees of freedom. This test had the hypothesis of H_0 where the model used followed the data, and H_1 where the model used did not follow the data. The results of Hosmer Lemeshow will reject H_0 if $P - value < \alpha$.

2.1.4. Coefficient Interpretation

Coefficient interpretation for the logistic regression model can be performed by seeing its odds ratio. Odds were defined as follows:

$$Odds = \frac{\pi_i}{1 - \pi_i} \tag{6}$$

where π_i stated success distribution (when $Y = 1$) and $1 - \pi_i$ stated failure distribution (when $Y = 0$). The odds ratio is the comparison of the odds value of two individuals. β_i parameter was defined as the change of logit function caused by a change in one unit of the $i-$th, yang disebut log odds, explanatory variable, which was called log odds [11], denoted as follows:

$$L_j(x_i) - L_j(x_{i+1}) = logit \frac{P(Y \leq j \mid x_i)/P(Y > j|x_i)}{P(Y \leq j \mid x_{i+1})/P(Y > j|x_{i+1})} \tag{7}$$

Thus, the estimator for the odds ratio was obtained as follows:

$$\widehat{OR} = \exp[\beta_i(x_i, x_{i+1})] \tag{8}$$

The odds ratio for the categorical explanatory variable, if it was more than 1, it was assumed to be m. Thus, it can be stated that odds when x_i were greater m-times than

odds when x_{i+1}. The odds ratio for continuous explanatory variable, when x increases by 1 unit, odds increased by $\exp[\beta_i\,(x_i, x_{i+1})]$ times than before.

2.2. Random Forest Ordinal

Random forest ordinal is a classification method used for categorical data and has stages in its categories. Based on [16], a statistical model for data with the ordinal response, such as proportional odds, has been investigated before, but this model has some weaknesses, including the model depends on certain assumptions that must be fulfilled. Moreover, the classic random forest introduced by He, et al. [17] also has some problems because it ignores the level information of the response variable. The version of random forest based on a unified framework for conditional inference, which provides unbiased selections of variables when looking for optimal split. The version of this random forest becomes an instrument that can be used for the data with the ordinal response because it provides the possibility to consider level information in the response variable.

Identification of explanatory variable correlated with response variable in random forest ordinal can be determined through Variable Importance Measures (VIMs). According to [16], three types of VIMs can be used in ordinal response, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Ranked Probability Score (RPS), which all of them consider level information in the response variable.

2.3. Model Evaluation

The best model evaluation used the accuracy of classification results from training data and test data. The results of classification from modeling can be measured its accuracy by observing the confusion matrix [18]. The accuracy value from the results of the classification can be seen in Table 2.

Table 2. Confusion matrix

Predict	Actual		
	No Diabetes	Prediabetes	Diabetes
No Diabetes	a	b	c
Prediabetes	d	e	f
Diabetes	g	h	i

From the confusion matrix in Table 2, the accuracy value can be calculated Accuracy $= \frac{a+e+i}{n}$ where n stated the number of observations.

3. Results and Discussions

3.1. Descriptive Analysis and Preliminary Data Exploration

Descriptive analysis was carried out on the response variables used in this study based on the data used by several criteria which are visually presented in Figure 1. Analysis factors causing diabetes mellitus, where the response variable was the diabetes status of respondents, was dominated by respondents who did not have diabetes with a proportion of 84%, while respondents with pre-diabetes and diabetes had a proportion of 14% and 2%, respectively as shown in Figure 1.

Furthermore, high blood pressure (BP) and cholesterol (HC) variables can be seen in Figure 2, where the highest respondents with non-BP and non-HC were in the

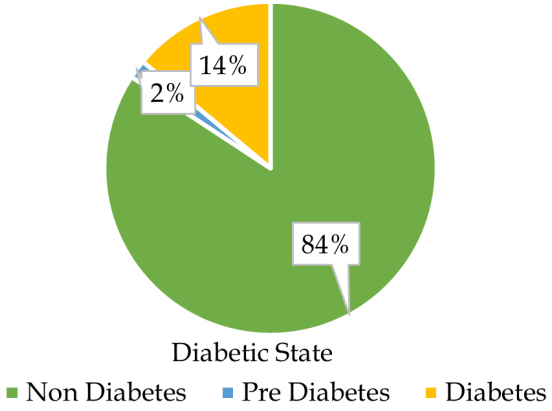

Figure 1. Proportion of diabetic state

non-diabetes category. Moreover, in the pre-diabetes category, the proportion decreased and the lowest was in the diabetes category. Likewise, the highest category of BP and HC was in the diabetes category. In the pre-diabetes category, the proportion increased and the lowest was in the non-diabetes category. This increases the assumption that BP and HC variables had a significant influence on diabetes status.

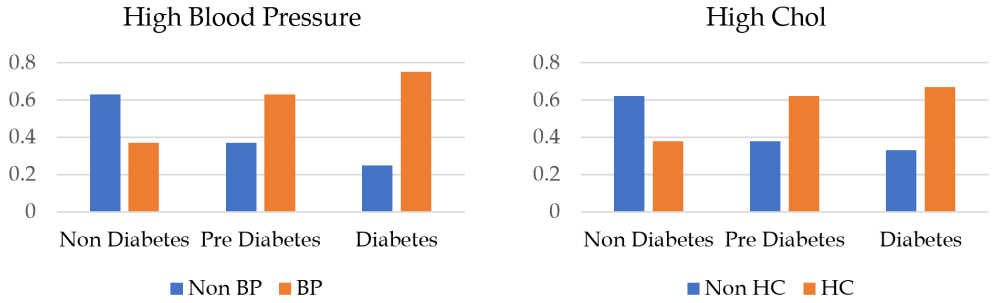

Figure 2. Correlation between diabetes status and High Blood Pressure and HighChol

Character from the data of body mass index (BMI) and age variables was explained using a boxplot in Figure 3.

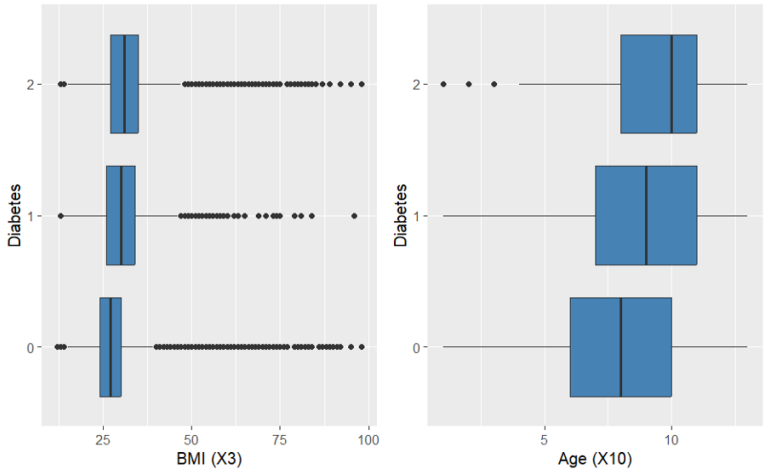

Figure 3. Correlation between diabetes status and BMI and age

It can be seen that the lowest mean in the boxplot of each variable was in the non-diabetes category. In the pre-diabetes category, the more it moved to the right the more it showed an increase and the highest was in the diabetes category. This increases the assumption that BMI and the variables had a significant influence on diabetes status. Furthermore, for Smoker and Gender variables can be seen in Figure 4. The graph had the same pattern as Figure 2, but the increase and decrease in diabetes status were not significant. This showed the possibility that Smoker and Gender variables had an influence on diabetes status although not as significant as 4 variables before.

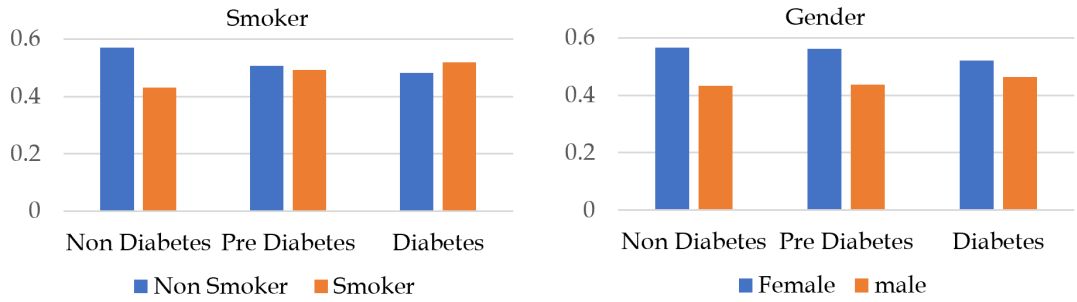

Figure 4. Correlation between diabetes status and smoker and sex

Fruit Consumption, Veggies Consumption, and Physical Activity in Figure 5 showed the opposite pattern to Figure 2 and Figure 4. The same case also can be seen in the Education variable, where category 1 to category 5 in Education had an increase in the pre-diabetes category and diabetes category even though not significant. Likewise, category 6 (Post Graduate) had a drastic decrease from non-diabetes to pre-diabetes and a slight decrease in the diabetes category. This showed the possibility that Fruit Consumption, Veggies Consumption, Physical Activity, and Education had a negative influence on diabetes status.

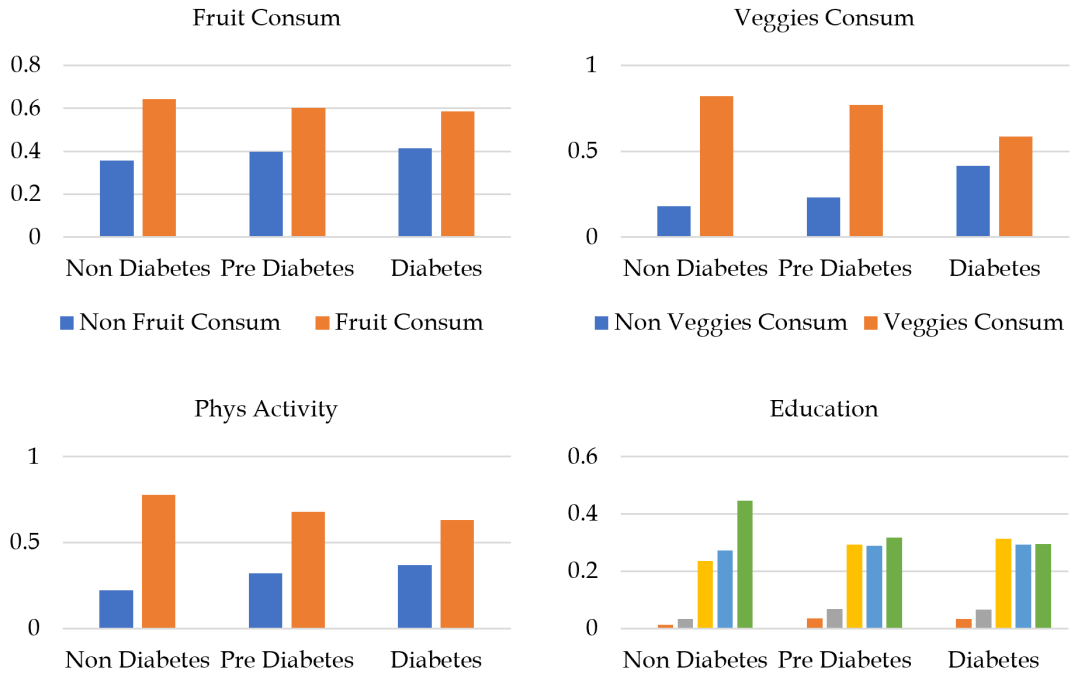

Figure 5. Correlation between diabetes status and fruit consumption, veggies, physical activity, and education

The Heavy Alcohol Consumption variable in Figure 6 showed either non-heavy alcohol consumption or heavy alcohol consumption graphs, in which both of them had the same proportion in each category of diabetes status.

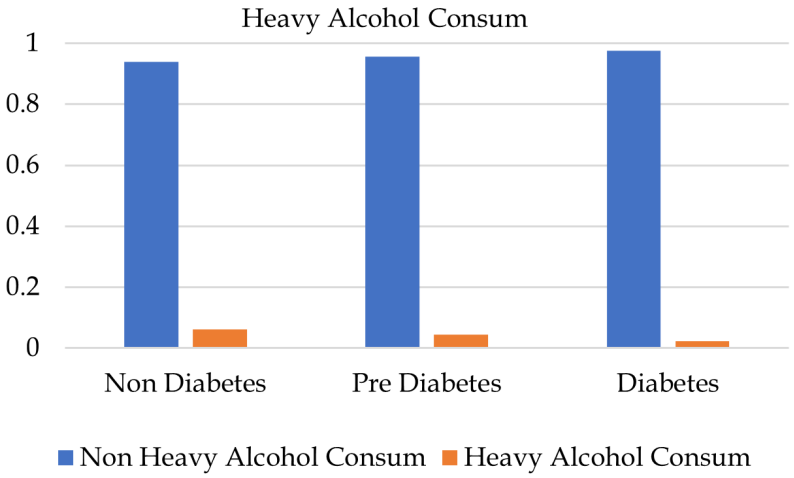

Figure 6. Correlation between diabetes status and alcohol consumption

This showed the possibility that Heavy Alcohol Consumption had the lowest influence over 10 other variables. After that, an assumption test was conducted in this exploration using ordinal logistic regression and random forest ordinal.

3.2. Ordinal Logistic Regression

3.2.1. Ordinal Logistic Regression Modeling

The model generated by ordinal logistic regression was obtained using training data. Model estimation used the Maximum Likelihood method. The results of the estimation of model parameters were presented in Table 3.

Table 3. Estimation of ordinal logistic regression model parameters

Variable	Coefficient $\widehat{\beta}_j$	Standard Eror
Explanatory variables		
BP	0.85637	0.075247
HighChol	0.78242	0.070655
BMI	0.06530	0.004656
Smoker	0.19616	0.067804
PhysActivity	-0.23098	0.073149
Fruits	-0.15394	0.070551
Veggies	0.02914	0.083620
Sex	0.11068	0.067725
HvyAlcoholConsum	-0.84779	0.193281
Age	0.13668	0.013655
Education	-0.22141	0.031926
Intercepts		
No Diabetes | Prediabetes	4.5026	0.2753
Prediabetes | Diabetes	4.6730	0.2758

The logit model equation formed based on Table 3 is as follows:

$$\begin{aligned} Logit\left[P\left(Y \leq 1 \mid X_j\right)\right] = 4.5026 + 0.85637X_1 + 0.78242X_2 + \\ 0.06530X_3 + 0.19616X_4 - 0.23098X_5 - \\ 0.15394X_6 + 0.02914X_7 + 0.11068X_8 - \\ 0.84779X_9 + 0.13668X_{10} - 0.22141X_{11} \end{aligned} \tag{9}$$

$$\begin{aligned} Logit\left[P\left(Y \leq 3 \mid X_j\right)\right] = 4.6730 + 0.85637X_1 + 0.78242X_2 + \\ 0.06530X_3 + 0.19616X_4 - 0.23098X_5 - \\ 0.15394X_6 + 0.02914X_7 + 0.11068X_8 - \\ 0.84779X_9 + 0.13668X_{10} - 0.22141X_{11} \end{aligned} \tag{10}$$

3.2.2. Parameter Testing

Parameter testing was divided into simultaneous testing and partial testing. Simultaneous testing aims to find the influence of the independent variable jointly on the response variable using Likelihood Ratio Test (G^2). After the analysis was conducted G^2 of 1207.555 was obtained, which was compared to $\chi^2_{(0.05,11)} = 19.675$, Since $G^2 > \chi^2_{(0.05,11)}$ then the decision to reject H0 means that at least there was one independent variable that had an influence on the response variable. Then the model was tested by partial testing.

Partial testing aims to find the influence of each independent variable on the response variable using the Wald test. The results of partial testing were presented in Table 4.

Table 4. Testing the parameters of the ordinal logistic regression model

Variable	Coefficient $\widehat{\beta}_j$	Standard Error	Wald's Value	p-value
Explanatory Variable				
BP	0.85637	0.075247	11.3808460	0.00000
HighChol	0.78242	0.070655	11.0737735	0.00000
BMI	0.06530	0.004656	14.0258797	0.00000
Smoker	0.19616	0.067804	2.8930264	0.00382
PhysActivity	-0.23098	0.073149	-3.1576606	0.00159
Fruits	-0.15394	0.070551	-2.1819587	0.02911
Veggies	0.02914	0.083620	0.3484236	0.72752
HvyAlcoholConsum	0.11068	0.067725	-4.3862870	0.00001
Sex	-0.84779	0.193281	1.6341882	0.10222
Age	0.13668	0.013655	10.0092025	0.00000
Education	-0.22141	0.031926	-6.9350708	0.00000
Intercepts				
No Diabetes | Prediabetes	4.5026	0.2753	16.3524544	0.00000
Prediabetes | Diabetes	4.6730	0.2758	16.9422859	0.00000

According to the table above, BP, HighChol, BMI, Smoker, Physical Activity, Fruits, Heavy Alcohol Consumption, Age, and Education had a p-value $< \alpha$ of 0.05 so the decision to reject H0 means that the nine variables had a significant influence on diabetes status. Meanwhile, Vegies and Gender variables had p-value $> \alpha$ so the

decision to reject H0 so the failed to reject H0 means both variables did not have a significant influence on diabetes status. Since there were non-significant variables in the model, then both variables were eliminated from the model. After elimination, re-testing was performed and obtained results in Table 5.

Table 5. Testing the parameters of the ordinal logistic regression model after elimination

Variable	Coefficient $\widehat{\beta}_j$	Standard Eror	Wald's Value	p-value
Explanatory Variable				
BP	0.86086025	0.075131663	11.458022	0.00000
HighChol	0.78201300	0.070648707	11.069035	0.00000
BMI	0.06520892	0.004648871	14.026831	0.00000
Smoker	0.21356133	0.066954827	3.189633	0.00142
PhysActivity	-0.22024614	0.072372335	-3.043237	0.00234
Fruits	-0.15805418	0.068492653	-2.307608	0.02102
HvyAlcoholConsum	-0.84745638	0.193148275	-4.387595	0.00001
Age	0.13543909	0.013634127	9.933829	0.00000
Education	-0.21626960	0.031700973	-6.822175	0.00000
Intercepts				
No Diabetes | Pre Diabetes	4.45611768	0.271718851	16.399737	0.00000
Pre Diabetes | Diabetes	4.62641509	0.272187633	16.997154	0.00000

According to the Table 5 above, all variables had p-value $< \alpha$ of 0.05 so it was a decision to reject H0, which means that after elimination, the other nine variables still had a significant influence on diabetes status.

3.2.3. Model Fit Testing

Model fit testing or Goodness of fit aims to find out whether the model used was suitable or not with data observed using Hosmer and Lemeshow Tests. According to the results of model fit testing, it was obtained p-value of 0.08555. P-value $>\alpha(0.05)$ showed that the model used was suitable for the data observed.

3.2.4. Odds Ratio and Interpretation

The results of the Odds Ratio and coefficient calculation are presented in Table 6. Thus, the equation of ordinal logistic regression was obtained as equation (11) and equation (12).

$$\begin{aligned} Logit\left[P\left(Y \leq 1 \mid X_j\right)\right] = {} & 4.45612 + 0.86095X_1 + 0.78201X_2 + 0.06521X_3 + \\ & 0.21356X_4 - 0.22025X_5 - 0.15805X_6 - 0.84746X_7 + \\ & 0.13544X_8 - 0.21627X_9 \end{aligned} \tag{11}$$

$$\begin{aligned} Logit\left[P\left(Y \leq 3 \mid X_j\right)\right] = {} & 4.62642 + 0.86095X_1 + 0.78201X_2 + 0.06521X_3 + \\ & 0.21356X_4 - 0.22025X_5 - 0.15805X_6 - 0.84746X_7 + \\ & 0.13544X_8 - 0.21627X_9 \end{aligned} \tag{12}$$

Table 6. Coefficient and odds ratio

Variable	Coefficient $\widehat{\beta}_j$	Odds Ratio
Explanatory Variable		
BP	0.86086025	2.3651945
HighChol	0.78201300	2.1858680
BMI	0.06520892	1.0673820
Smoker	0.21356133	1.2380794
PhysActivity	-0.22024614	0.8023213
Fruits	-0.15805418	0.8538035
HvyAlcoholConsum	-0.84745638	0.4285035
Age	0.13543909	1.1450394
Education	-0.21626960	0.8055181
Intercepts		
No Diabetes | Pre Diabetes	4.45611768	86.1523874
Pre Diabetes | Diabetes	4.62641509	102.1472180

From the equation (11)-(12), and the odds ratio in the Table 6, it can be interpreted as follows:

1. Intercept diabetes status in the non-diabetes category had an odds ratio of 86.1523874, which means that diabetes status distribution in the non-diabetes category was 86.1523874 times compared to the distribution to be pre-diabetes category with the assumption of the constant independent variable.
2. Intercept diabetes status of the diabetes category had an odds ratio of 102.1472180, which means that diabetes status distribution in the diabetes category was 102.1472180 times compared to the distribution being pre-diabetes category with the assumption of the constant independent variable.
3. Blood Pressure variable had an odds ratio of 2.3651945, which means that every person who had hypertension will increase the odds of diabetes status in the diabetes category of 2.3651945 compared to odds in pre-diabetes and non-diabetes categories.
4. High Chol variable had an odds ratio of 2.1858680, which means that every person who had high cholesterol will increase the odds of diabetes status in the diabetes category of 2.1858680 compared to odds in the pre-diabetes and non-diabetes categories.
5. BMI variable had an odds ratio of 1.0673820. which means that every 1 unit increase in BMI will increase the odds of diabetes status in the diabetes category of 1.0673820 compared to odds in pre-diabetes and non-diabetes categories.
6. Smoker variable had an odds ratio of 1.2380794, which means that if someone smokes more than 100 cigarettes in a lifetime, it will increase the odds of diabetes status in the diabetes category of 1.2380794 compared to the odds in pre-diabetes and non-diabetes categories.
7. Physical Activity variable had an odds ratio of 0.8023213, which means that if someone has physical activities besides working within 30 days before the examination, it will decrease the odds of diabetes status in the diabetes category of 0.8023213 compared to odds in pre-diabetes and non-diabetes categories.
8. Fruits variable had an odds ratio of 0.8538035, which means that if someone consumes fruits every day, it will decrease the odds of diabetes status in the diabetes category of 0.8538035 compared to the odds in pre-diabetes and

non-diabetes categories.

9. Heavy Alcohol Consumption variable had an odds ratio of 0.4285035, but it cannot be interpreted that if someone consumes alcohol more than 7 glasses per week, it will decrease the odds of diabetes status in the diabetes category of 0.4285035 compared to odds in pre-diabetes and non-diabetes categories. However, it was illogical, that the content of alcohol contrarily will increase the chance of a person having diabetes.
10. Age variable had an odds ratio of 1.1450394, which means that every 1 unit increase in the age group will increase the odds of diabetes status in the diabetes category of 1.1450394 compared to odds in pre-diabetes and non-diabetes categories.
11. Education variable had an odds ratio of 0.8055181, which means that every 1 unit increase in education will decrease the odds of diabetes status in the diabetes category of 0.8055181 compared to odds in the pre-diabetes and non-diabetes categories.

3.3. Random Forest Ordinal

Model formation in the classification of random forest ordinal used 11 explanatory variables, which were estimated to have an influence on diabetes status. The number of trees (k) used to build a random forest ordinal model in this study was 500 trees and the number of sorting variables used was three variables. Before determining the number of trees, 50, 100, 500, to 1000 trees were observed previously and the evaluation average was calculated using cross-validation to determine optimal trees used in modeling random forest ordinal. Cross-validation was carried out by repeating 100 times and the average value was obtained.

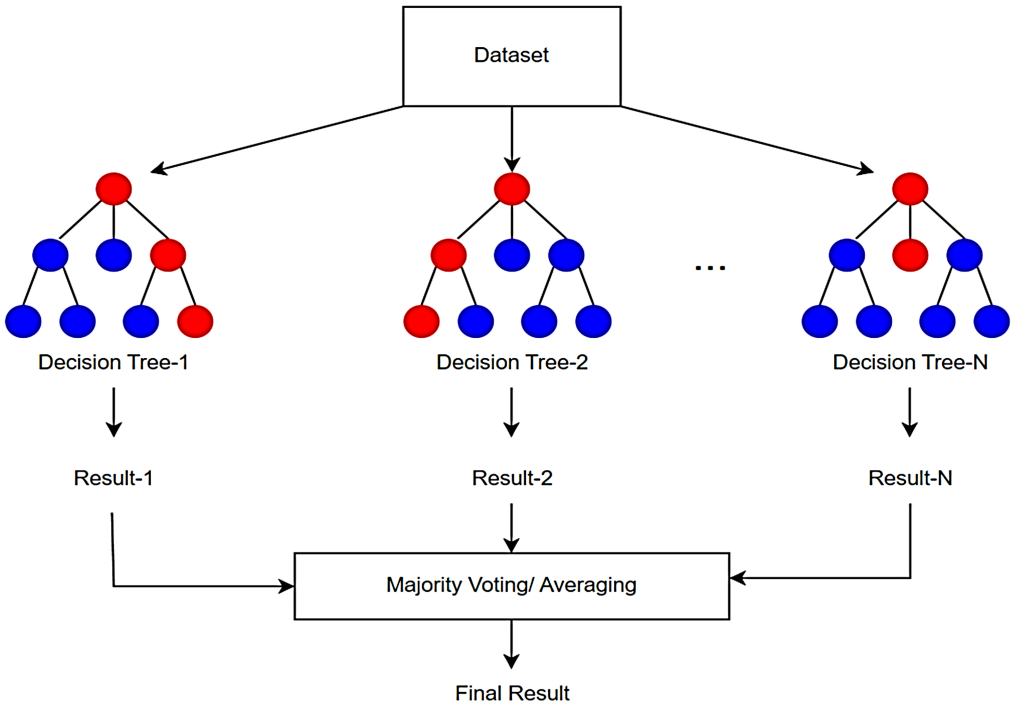

Figure 7. How the Random Forest Algorithm Works where N=500

The important measure of the explanatory variable used in this study was VIMs-RPS

because it is in accordance with the previous study, which stated that the best measurement of the characteristic variable for the ordinal response was VIMs-RPS Figure 7 showed the illustration of 4 explanatory variables with the highest level of variable importance with 500 formed trees and sorting variable.

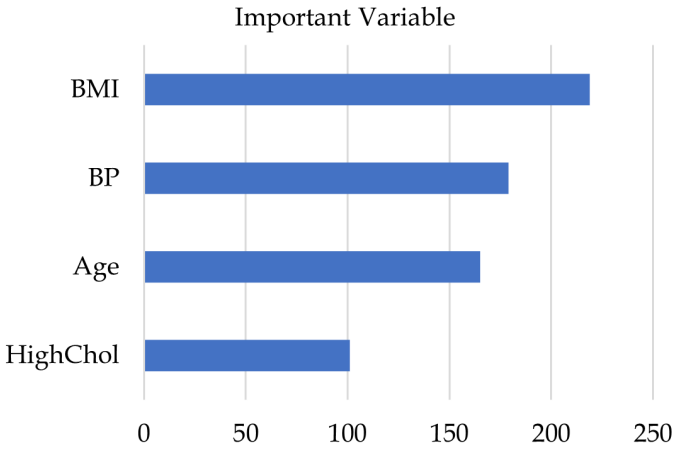

Figure 8. Important variables according to the ordinal random forest method

It can be seen that there were ten explanatory variables with a higher importance value of explanatory variable than other explanatory variables. The explanatory variables were BMI, BP, Age, and HighChol as estimated in data exploration before.

3.4. Comparing the Results of Model Evaluation

Ordinal logistic regression and random forest formed before were evaluated by testing model accuracy for the testing data using a confusion matrix. The confusion matrix of training data and test data was presented in Table 7 and Table 8.

Table 7. Confusion matrix ordinal logistic regression model

Prediction	Actual		
	No Diabetes	Prediabetes	Diabetes
No Diabetes	1690	36	258
Prediabetes	0	0	0
Diabetes	19	1	25

Table 8. Confusion matrix model random forest ordinal

Prediction	Actual		
	No Diabetes	Prediabetes	Diabetes
No Diabetes	1596	73	40
Prediabetes	32	3	2
Diabetes	221	35	26

According to the results of the confusion matrix, it was obtained accuracy value using ordinal logistic regression of 84.52% and an accuracy value using a random forest of 80.13% so that both methods used were categorized as suitable for prediction. Ordinal logistic regression generates a higher accuracy value than random forest ordinal. Ordinal logistic regression has higher accuracy but it cannot detect any respondent in the pre-diabetes category.

4. Conclusion

The best model obtained in this study is ordinal logistic regression because it generates a higher accuracy value than random forest ordinal. When a person has hypertension, has cholesterol and smokes the likelihood that he will have diabetes will be higher than those who do not have hypertension, have no cholesterol and do not smoke. In addition, the possibility of diabetes also increases with age and body mass index. On the other hand, when a person regularly does physical activity outside of work, eats fruit and has a high level of education, his chances of developing diabetes will decrease. The four most important variables causing diabetes are body mass index, hypertension, age, and cholesterol. Future study is expected to perform imbalance in the data so that the prediction accuracy of the pre-diabetes category can increase.

References

[1] C. M. Ridhani, N. A. W. Putri, R. A. Fitriani, I. A. A. Rahmayati, and A. G. Rizka, "Analysis of trigger factors for diabetes based on cdc data," *Jurnal Teknodik Pustekkom Kemdikbud*, vol. 2022, pp. 1–7, 2022, doi: 10.13140/RG.2.2.14821.27362.

[2] Y. Olviani and D. Novita, "Family support on blood sugar control compliance with diabetes mellitus patients during the covid-19 pandemic," *Jurnal Keperawatan Suaka Insan (JKSI)*, vol. 7, no. 2, pp. 178–183, 2022.

[3] A. Siahaan, E. I. Marpaung, and J. Pandaleke, "Tatalaksana geriatri dengan diabetes melitus," *Jurnal Medik dan Rehabilitasi*, vol. 4, no. 3, pp. 1–7, 2022.

[4] R. Kemenkes, "Infodatin pusat data dan informasi kementerian kesehatan republik indonesia," 2020, [Online] available at https://www.kemkes.go.id/downloads/resources/download/pusdatin/infodatin/Infodatin%202020%20Diabetes%20Melitus.pdf.

[5] T. D. Cahyono and O. S. Purwanti, "Hubungan antara lama menderita diabetes dengan nilai ankle brachial index," *Jurnal Berita Ilmu Keperawatan*, vol. 12, no. 2, pp. 65–71, 2019, doi: 10.23917/bik.v12i2.9803.

[6] R. Anggraini, "Korelasi kadar kolesterol dengan kejadian diabetes mellitus tipe 2 pada laki-laki," *Medical and Health Science Journal*, vol. 2, no. 2, pp. 55–60, 2018, doi: 10.33086/mhsj.v2i2.588.

[7] I. D. G. I. P. Putra, I. A. P. Wirawati, and N. N. Mahartini, "Hubungan kadar gula darah dengan hipertensi pada pasien diabetes mellitus tipe 2 di rsup sanglah," *Intisari Sains Medis*, vol. 10, no. 3, pp. 797–800, 2019, doi: 10.15562/ism.v10i3.482.

[8] Z. I. Nisa, A. M. Soleh, and H. Wijayanto, "Identifikasi faktor-faktor yang memengaruhi prestasi mahasiswa menggunakan regresi logistik ordinal dan random forest ordinal," *Xplore: Journal of Statistics*, vol. 10, no. 1, pp. 88–101, 2021, doi: 10.29244/xplore.v10i1.465.

[9] L. B. C. Tanujaya, B. Susanto, and A. Saragih, "The comparison of logistic regression methods and random forest for spotify audio mode featurre classification," *Indonesian Journal of Data and Science*, vol. 1, no. 3, pp. 68–78, 2020, doi: 10.33096/ijodas.v1i3.16.

[10] R. Susetyoko, W. Yuwono, E. Purwantini, and N. Ramadijanti, "Perbandingan metode random forest, regresi logistik, naïve bayes, dan multilayer perceptron pada klasifikasi uang kuliah tunggal (ukt)," *Jurnal Infomedia*, vol. 7, no. 1, pp. 8–16, 2022, doi: 10.30811/jim.v7i1.2916.

[11] E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *Journal of Data Analysis and Information Processing*, vol. 07, no. 04, pp. 190–207, 2019, doi: 10.4236/jdaip.2019.74012.

[12] T. J. Smith, D. A. Walker, and C. M. McKenna, "An exploration of link functions used in ordinal regression," *Journal of Modern Applied Statistical Methods*, vol. 18, no. 1, pp. 2–15, 2020, doi: 10.22237/jmasm/1556669640.

[13] Y. Paliling, M. Fathurahman, and S. Wahyuningsih, "Multinomial logistic regression to model the combination of phdi status and hdi status of districts/cities in kalimantan island," *Jurnal Matematika, Statistika dan Komputasi*, vol. 19, no. 3, pp. 460–472, 2023, doi: 10.20956/j.v19i3.22299.

[14] N. I. Mardini, L. Marlena, and E. Azhar, "Regresi logistik pada model problem based learning berbantu software cabri 3d," *Jurnal Mercumatika: Jurnal Penelitian Matematika dan Pendidikan Matematika*, vol. 4, no. 1, pp. 47–53, 2019.
[15] M. Díaz-Pérez, Ángel Carreño-Ortega, J.-A. Salinas-Andújar, and Ángel Jesús Callejón-Ferre, "Application of logistic regression models for the marketability of cucumber cultivars," *Agronomy*, vol. 9, no. 1, p. 17, 2019, doi: 10.3390/agronomy9010017.
[16] V. Syrgkanis and M. Zampetakis, "Estimation and inference with trees and forests in high dimensions," in *Proceedings of Machine Learning Research*, 2020, p. 125.
[17] Y.-X. He, S.-H. Lyu, and Y. Jiang, "Interpreting deep forest through feature contribution and mdi feature importance," *arXiv preprint*, 2023.
[18] T. Y. Yuniarty, Erfiani, Indahwati, A. Fitrianto, and K. N. K., "Regresi ordinal logit dan probit pada faktor kesejahteraan rumah tangga petani tanaman pangan di provinsi sulawesi tenggara," *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 2, pp. 313–325, 2022, doi: 10.21009/JSA.06216.