Universitas Brawijaya

Application of Logistic Regression in Machine Learning for Predicting Diabetes Patients

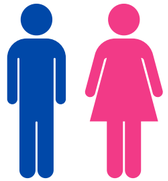
Zheryl Zabrina Hibatullah, S.Stat
(Consultant and Data Analyst)

Program Kuliah Praktisi Universitas Brawijaya 2025 | Sarjana Statistika |Analisis Data Kategorik

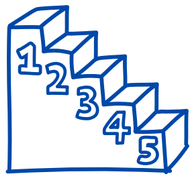
1. Introduction to Categorical Data

What is Categorical Data?

- Categorical Data = data grouped into categories (not numerical)

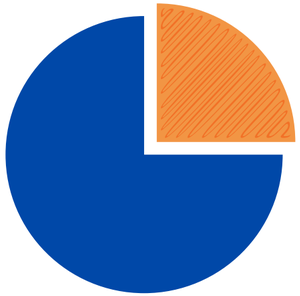


Nominal
e.g., Gender (Male/Female), Blood Type (A/B/AB/O)

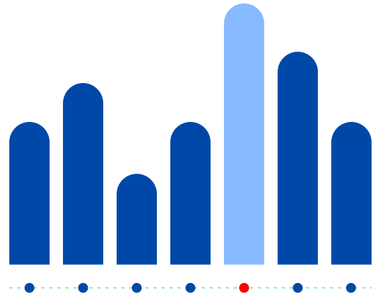


Ordinal
e.g., *Pain Level (Mild / Moderate / Severe)*

Visualization



Pie Chart



Bar Chart

2. What is Logistic Regression?

A statistical model used to:

🎯 To model the relationship between a categorical response variable (Y) and one or more predictor variables (X)

🔗 Model probabilities using the logit function

🧠 Interpret results with odds ratios

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

3. Why Logistic Regression?

	❌ Linear Regression	✅ Logistic Regression
🔢 Output Type	Continuous (can be <0 or >1)	Probability (bounded between 0 and 1)
📉 Problem	Predicts impossible probabilities	Predicts valid probabilities
📊 Assumptions	Requires normality & equal variance (homoscedasticity)	No strict normality or equal variance needed
❓ Interpretation	Difficult to interpret with categorical outcomes	Provides interpretable odds ratios
🔄 Use Case	Best for numeric outcomes	Best for binary / categorical outcomes

4. Assumtion in Logistic Regression

Binary Outcome (Y)

The dependent variable must be categorical
(e.g., 1 = Diabetes, 0 = No Diabetes)

Independent Observations

Each case/row is assumed to be independent of others

(⚠️ No repeated measurements or clustering)

No Multicollinearity

Predictors should not be strongly correlated
✅ Check VIF (VIF<10) or correlation matrix

5. Inference in Logistic Regression

Confidence Interval for β

$$\hat{\beta} \pm z_{\alpha/2} \cdot SE$$

Partial Significance (Wald Test)

Hypotheses:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

Test Statistic:

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Approx. follows standard normal distribution when sample size is large

Simultaneous Significance (Likelihood Ratio Test)

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \text{At least one } \beta_j \neq 0$$

Test Statistic:

$$G = -2(L_0 - L_p)$$

- L0: log-likelihood of model without predictors
- Lp: log-likelihood of model with predictors

📈 GGG follows Chi-Square distribution with df = number of predictors

6. Goodness of fit test

📊 Metric/Test	✅ Interpretation / Criteria
Deviance	Lower residual deviance → better model fit
Likelihood Ratio Test	Significant p-value (e.g., < 0.05) → model explains variation
Hosmer–Lemeshow Test	p-value > 0.05 → good fit (fail to reject null = model fits data)
Pseudo R^2 (e.g., McFadden)	McFadden's R^2 > 0.2 = decent fit (but no strict threshold)

6. Model Evaluation

Confusion Matrix

	Actual: Positive	Actual: Negative
Predicted: Positive	✅ TP (True Positive)	❌ FP (False Positive)
Predicted: Negative	❌ FN (False Negative)	✅ TN (True Negative)

Performance Metrics

Sensitivity

$$\frac{TP}{TP+FN}$$

True Positive Rate (Recall)

Specificity

$$\frac{TN}{TN+FP}$$

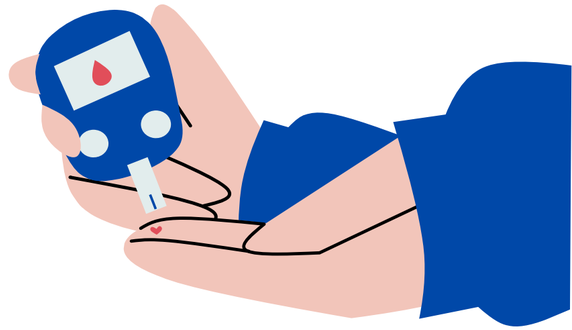
True Negative Rate

Accuracy

$$\frac{TP+TN}{TP+TN+FP+FN}$$

Overall correct classification

Case Study : Diabetes Risk Prediction



- Diabetes is a serious, chronic illness affecting 34.2M Americans, with 1 in 5 undiagnosed
- Early prediction can save lives and reduce healthcare costs.
- This project aims to predict diabetes risk using health survey data from the CDC's BRFSS 2015 — the largest ongoing U.S. health survey.

	Details
👥 Sample Size	253,680 respondents
🧪 Target Variable	Diabetes: 0 = No, 1 = Diabetes
📈 Features	21 health indicators
📞 Source	BRFSS (Behavioral Risk Factor Surveillance System), CDC

Data Preprocessing

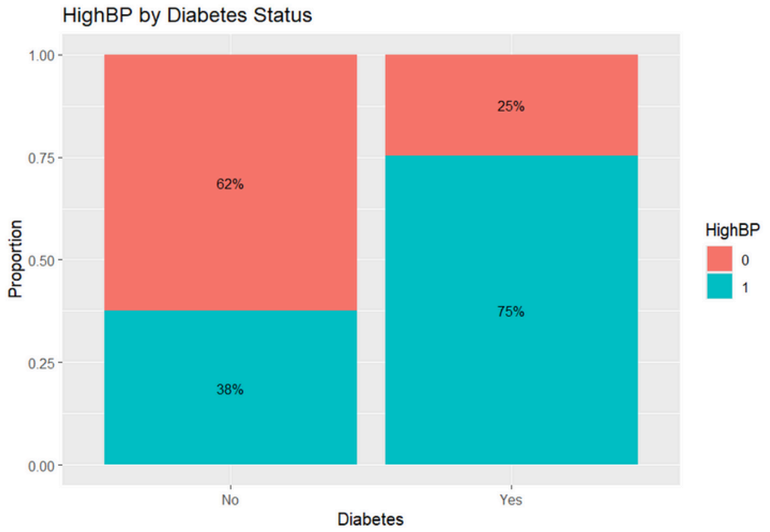
Exploratory Data Analysis (EDA)

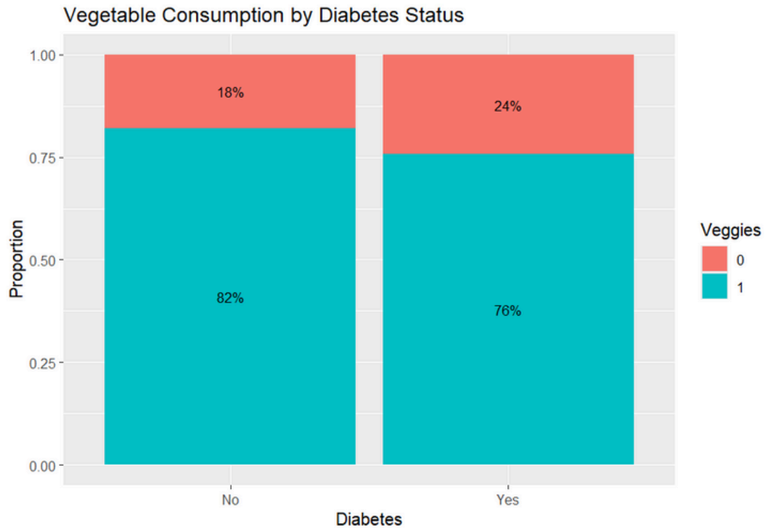
Model Building: Logistic Regression

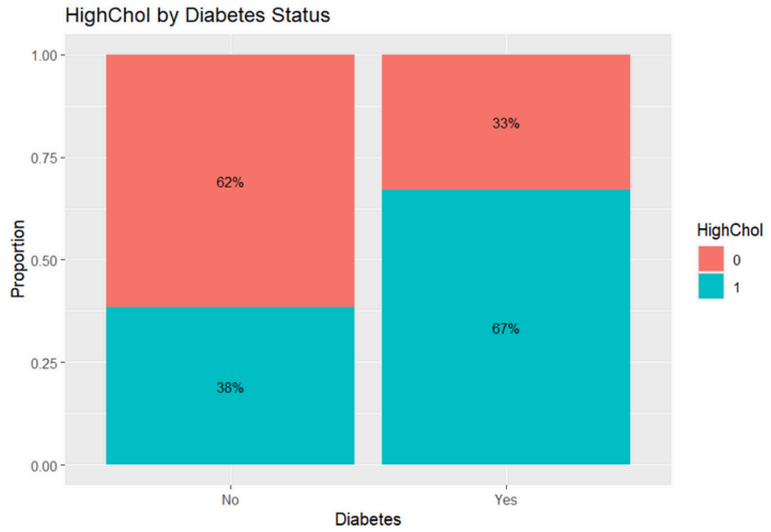
Model Evaluation

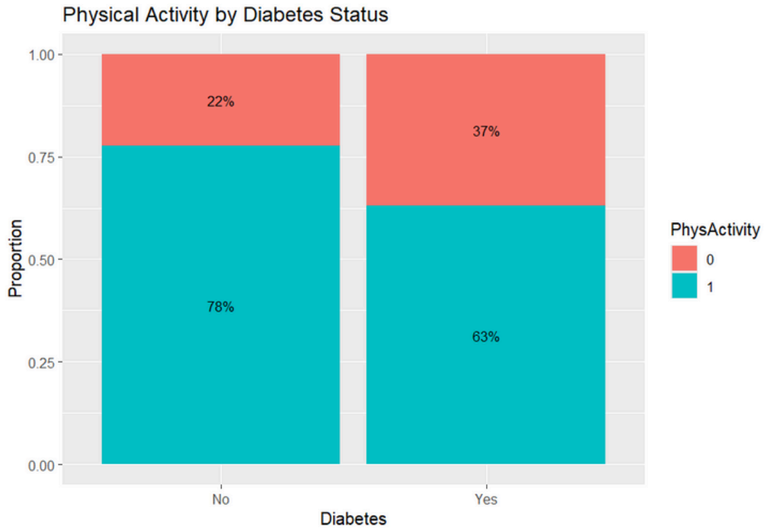
Inference & Interpretation

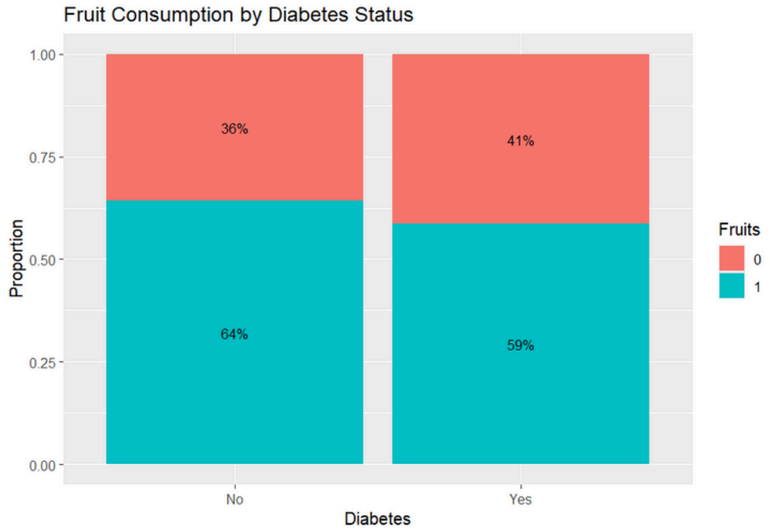
Exploratory Data Analysis (EDA)

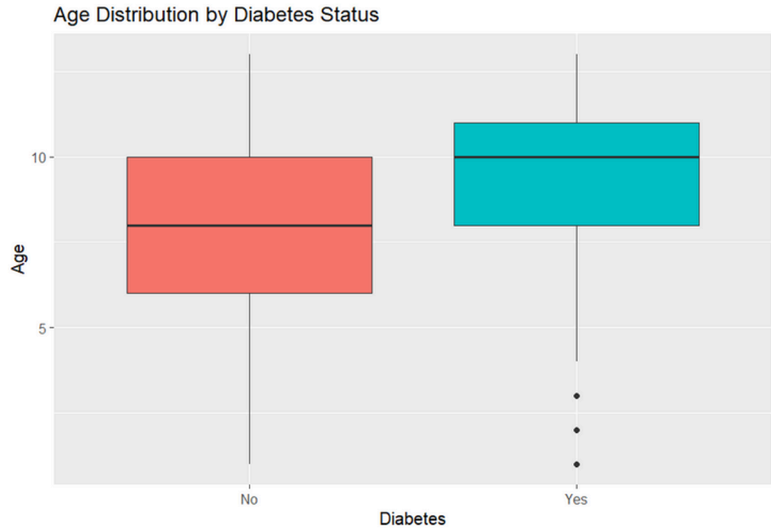












Check No Multicollinearity

```
> # Multicollinearity Check: VIF
> vif(model)
      HighBP     HighChol PhysActivity          Age       Fruits      Veggies
    1.109727     1.061752     1.041495     1.103111     1.098789     1.086687
```

- All VIF values are close to 1 and well below 5.
- Interpretation: There is no multicollinearity issue among the predictors.
- The independent variables are not strongly correlated with each other.

Summary Logistic Regression Model

```
Call:
glm(formula = Diabetes_binary ~ HighBP + HighChol + PhysActivity +
    Age + Fruits + Veggies, family = binomial, data = train_data)

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -3.223075   0.029659 -108.670   <2e-16 ***
HighBP        1.190664   0.015636   76.151   <2e-16 ***
HighChol      0.723199   0.014447   50.059   <2e-16 ***
PhysActivity -0.452869   0.014635  -30.943   <2e-16 ***
Age           0.100364   0.002701   37.164   <2e-16 ***
Fruits       -0.135328   0.014474   -9.350   <2e-16 ***
Veggies      -0.165668   0.016816   -9.852   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 163886  on 202943  degrees of freedom
Residual deviance: 143479  on 202937  degrees of freedom
AIC: 143493

Number of Fisher Scoring iterations: 5
```

Logistic Regression Equation

The fitted logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = -3.223 + 1.191 \cdot \text{HighBP} + 0.723 \cdot \text{HighChol} - 0.453 \cdot \text{PhysActivity} + 0.100 \cdot \text{Age} - 0.135 \cdot \text{Fruits} - 0.166 \cdot \text{Veggies}$$

Summary Logistic Regression Model

Odds Ratio

Variable	OR	Interpretation
HighBP	3.29	People with high blood pressure are 3.29 times more likely to have diabetes.
HighChol	2.06	Those with high cholesterol are 2.06 times more likely to have diabetes.
PhysActivity	0.64	Physically active individuals have 36% lower odds of diabetes.
Age	1.11	Each unit increase in age increases the odds by 11%.
Fruits	0.87	Fruit consumption is linked to 13% lower odds of diabetes.
Veggies	0.85	Vegetable consumption is associated with 15% lower odds of diabetes.

R Square

McFadden's $R^2 = 0.125$

Interpretation: The model explains about 12.5% of the variability in diabetes status, which is acceptable for logistic models in public health data.

Summary Logistic Regression Model

Partial Significance (Wald Test)

Hypotheses

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$ for i = 1,2,3,4,5,6

Significance Level

$\alpha = 0.05$

p-values

All p-values < 2e-16

Decision

Since $p < \alpha$, reject H_0

Conclusion

Each predictor individually has a significant effect on the likelihood of heart disease at the 5% level.

Simultaneous Significance (Likelihood Ratio Test)

Hypotheses

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

H_1: At least one $\beta_i \neq 0$ for i = 1 to 6

Test Statistic & Critical Value

- $G^2 = 20{,}406.4$
- $\chi^2(0.05, 6) = 12.59$

Decision

Since $G^2 > \chi^2$, reject H_0

Interpretation

At the 5% significance level, the predictors — high blood pressure, high cholesterol, physical activity, age, fruit consumption, and vegetable consumption — jointly have a significant effect on diabetes status.

Model Evaluation

Confusion Matrix

	Actual: Positive	Actual: Negative
Predicted: Positive	43,669	7,067
Predicted: Negative	0	0

📊 Model Evaluation Metrics

Accuracy: 86.1%

→ The model correctly classified 86.1% of all cases.

Sensitivity (Recall for positive class): 0%

→ The model failed to identify any individuals with diabetes (class 1).

Specificity: 100%

→ The model perfectly identified all non-diabetic individuals.

Thanks