

ABSTRAK

Regresi logistik adalah metode pemodelan statistis untuk variabel terikat kategorik (umumnya biner) yang bertujuan mengestimasi probabilitas terjadinya suatu peristiwa, berbeda dengan regresi klasik yang memodelkan nilai kontinu. Perbedaan utama dengan regresi linear terletak pada bentuk model, di mana regresi logistik memodelkan hubungan linear antara variabel bebas dan *log-odds* dari probabilitas kejadian, sehingga probabilitas yang dihasilkan selalu berada pada rentang 0 hingga 1. Koefisien regresi logistik diinterpretasikan sebagai *odds ratio* dan diestimasi menggunakan pendekatan *maximum likelihood* tanpa memerlukan asumsi normalitas dan homoskedastisitas galat, serta dievaluasi berdasarkan kecocokan dan kemampuan klasifikasi model.

Tugas akhir ini menerapkan model regresi logistik untuk mengidentifikasi faktor-faktor yang memengaruhi kejadian hipertensi berdasarkan hasil pengukuran. Faktor-faktor yang digunakan adalah faktor sosiodemografi berdasarkan data Survei Kesehatan Indonesia (SKI) 2023 dengan sampel sebanyak 638.178. Karakteristik data kategori yang melekat pada variabel-variabel bebas menuntut penggunaan variabel dummy dalam proses pemodelan. Proses pencarian model terbaik melibatkan beberapa teknik seperti penyeimbangan data, validasi silang, dan seleksi variabel. Hasil penelitian menghasilkan model dengan nilai akurasi 63,91% dan nilai sensitivitas 70.45% yang menunjukkan kemampuan model cukup baik dalam mengidentifikasi penderita hipertensi. Hasil penelitian menunjukkan karakteristik wilayah, jenis kelamin, status pernikahan, umur, status pekerjaan, serta kepemilikan Kartu Keluarga Sejahtera (KKS) berpengaruh terhadap risiko hipertensi.

Kata kunci: Regresi logistik, Validasi silang *k*-lipat, Hipertensi.

ABSTRACT

Logistic regression is a statistical modeling method for categorical (usually binary) dependent variables that aims to estimate the probability of an event occurring, unlike classical regression, which models continuous values. The main difference from linear regression lies in the model format. Logistic regression models a linear relationship between the independent variables and the log-odds of the event probability, so the resulting probability always ranges from 0 to 1. Logistic regression coefficients are interpreted as *odds ratios* and estimated using a maximum likelihood approach without requiring assumptions of normality and homoscedasticity of errors. They are evaluated based on the model's fit and classification ability.

This Undergraduate Thesis applies a logistic regression model to identify factors influencing the incidence of hypertension based on measurement results. The factors used are sociodemographic factors based on data from the 2023 Indonesian Health Survey (SKI) with a sample of 638,178. The categorical data characteristic of the independent variables requires the use of dummy variables in the modeling process. The process of finding the best model involves several techniques such as data balancing, cross-validation, and variable selection. The study produced a model with an accuracy of 63.91% and a sensitivity of 70.5%, indicating a good ability to identify hypertension sufferers. The study also showed that characteristics of rural areas, gender, marital status, age, employment status, and ownership of a Prosperous Family Card (KKS) influenced the risk of hypertension.

Keywords: Logistic Regression, *k*-Fold Cross-Validation, Hypertension.