INTISARI

Penerapan Seleksi Fitur untuk Meningkatkan Performa Klasifikasi pada *Dataset* Medis Pasien Gagal Jantung

Oleh
Leslie Anggraini
21/475917/PPA/06142

Gagal jantung merupakan penyakit kardiovaskular dikarenakan penurunan fungsi relaksasi atau kontraksi jantung. Penyakit ini mengakibatkan jantung tidak mampu memompa darah dengan cukup keseluruh tubuh, menyebabkan penumpukan cairan dalam tubuh yang meningkatkan risiko komplikasi serius dan berpotensi fatal pada kesehatan. Terdapat faktor-faktor risiko yang mempengaruhi pasien gagal jantung. Keseluruhan faktor tersebut perlu diidentifikasi untuk memahami stratifikasi risiko dengan mengetahui pasien yang berisiko tinggi terhadap kematian. Stratifikasi ini melibatkan pengelompokan atau klasifikasi pada pasien gagal jantung.

Penelitian ini menggunakan *dataset* medis pasien gagal jantung yang diekstraksi dari *The Medical Information Mart for Intensive Care* (MIMIC-III) yang berisi 1.177 data. Klasifikasi menggunakan keseluruhan fitur yang sangat banyak (*multivariate*) pada *dataset* ini dapat menghasilkan performa klasifikasi yang kurang optimal, terutama dengan spesifisitas yang rendah. Berdasarkan hal ini, diperlukan penerapan seleksi fitur yang mengurangi jumlah fitur untuk mengoptimalkan performa klasifikasi.

Terdapat beberapa tahapan yang diterapkan, dimulai dari pra-pemrosesan, seleksi fitur, membangun model klasifikasi, dan evaluasi model. Pra-pemrosesan mencakup dua proses, yaitu imputasi nilai untuk mengatasi data dengan nilai yang kosong dan normalisasi untuk menyeragamkan rentang nilai data. Seleksi fitur menggunakan beberapa metode, seperti *Pearson Correlation, Information Gain,* Chi-*Squared,* Ridge, *Forward Selection,* dan *Backward Elimination* sebagai perbandingan dalam memperoleh *subset* fitur yang paling relevan dan berkontribusi signifikan terhadap klasifikasi. Klasifikasi menggunakan metode *Logistic Regression* dan *Support Vector Machine* (SVM) untuk mencari performa yang lebih baik dalam mengklasifikasi pasien gagal jantung. Hasil penelitian ini menunjukkan bahwa penerapan seleksi fitur berhasil meningkatkan performa klasifikasi, khususnya dalam spesifisitas. Peningkatan performa ini meningkatkan akurasi sebesar 2%, presisi sebesar 3%, spesifisitas sebesar 55%, dan f1-score sebesar 2%.

Kata Kunci: Seleksi Fitur, *Correlation, Information Gain,* Chi-*Squared,* Ridge, *Forward Selection, Backward Elimination,* Klasifikasi, *Logistic Regression,* SVM, Gagal Jantung.

ABSTRACT

Implementing Feature Selection to Improve Classification Performance on The Medical Dataset of Heart Failure Patients

Created by
Leslie Anggraini
21/475917/PPA/06142

Heart failure is a cardiovascular disease caused by the heart's decreased relaxation or contraction function. This condition results in the heart's inability to pump blood sufficiently to the entire body, leading to fluid accumulation in the body that increases the risk of severe complications and potentially fatal health outcomes. There are risk factors that influence heart failure patients. Identifying these factors is crucial for understanding risk stratification and identifying patients at high mortality risk. Stratification involves grouping or classifying heart failure patients. This study utilizes a medical dataset of heart failure patients extracted from The Medical Information Mart for Intensive Care (MIMIC-III), contain 1,177 instances. Classification using this dataset's many features (multivariate) can result in suboptimal classification performance, particularly with low specificity. Therefore, feature selection is applied to reduce the number of features and optimize classification performance.

Several stages are used, from pre-processing, feature selection, classification model building, and evaluating the model. Pre-processing includes two processes: imputing values to address missing data and normalizing the data value range. Feature selection employs various methods, such as Pearson Correlation, Information Gain, Chi-Squared, Ridge, Forward Selection, and Backward Elimination, for comparison to obtain the most relevant and significantly contributing feature subset for classification. Classification uses Logistic Regression and Support Vector Machine (SVM) methods to seek better performance in classifying heart failure patients. The results of this study indicate that implementing feature selection improves classification performance, especially regarding specificity. The performance improvement leads to a 2% increase in accuracy, a 3% increase in precision, a 55% increase in specificity, and a 2% increase in the f1-score.

Keywords: Feature Selection, Correlation, Information Gain, Chi-Squared, Ridge, Forward Selection, Backward Elimination, Classification, Logistic Regression, SVM, Heart Failure.