Komparasi Kinerja Metode Random Forest Regression dengan Metode Support Vector Regression untuk Memprediksi Usia Biologis pada Data Pemeriksaan Medis = Comparison of the Performance of the Random Forest Regression Method with the Support Vector Regression Method for Predicting Biological Age on Medical Examination Data

Nadia Hartini Kusumawijaya, author

Abstrak

Penuaan adalah salah satu faktor utama resiko terjadinya penyakit dan kematian. Laju penuaan individu dengan usia kronologis yang sama terbukti bervariasi. Maka dari itu, muncul kebutuhan untuk alat pengukuran penuaan yang lebih akurat, robust, dan dapat diandalkan dibandingkan usia kronologis, yakni usia biologis. Pada penelitian ini, penulis membangun model menggunakan Metode Random Forest Regression (RF) dan Metode Support Vector Regression (SVR) untuk memprediksi umur biologis pada data pemeriksaan medis, menilai dan mengevaluasi hasil kinerjanya, serta melakukan komparasi kinerja kedua metode. Terkait metode yang digunakan, Metode RF adalah metode yang mengaplikasikan Teknik Ensemble Learning dengan cara menggabungkan beberapa decision tree untuk menghasilkan prediksi. Sedangkan, Metode SVR adalah metode yang berkerja dengan cara membangun hyperplane atau kumpulan hyperplane dalam ruang berdimensi tinggi yang dapat digunakan untuk regresi linier atau nonlinier. Dataset yang digunakan adalah data medis yang berasal dari Kementrian Kesehatan Republik Indonesia. Pada dataset dilakukan data preprocessing, yakni data diproses pada aspek missing values handling, encoding, dan outliers detection and outliers handling. Kemudian, dilakukan feature selection menggunakan Spearman’s Rank Correlation Coefficient. Setelah itu, dilakukan pembangunan model dengan Metode RF dan model dengan Metode SVR secara terpisah untuk masing - masing jenis kelamin. Terakhir, performa model dievaluasi dan dibandingkan kinerjanya menggunakan metrik evaluasi Root Mean Square Error (RMSE), Coefficient of Determination (R2), Adjusted R2, dan running time. Metode RF menggunakan hyperparameter terbaik {’max depth’: 15, ’n estimators’: 1150} untuk dataset pria, dan {’max depth’: 15, ’n estimators’: 1250} untuk dataset wanita. Sedangkan, Metode SVR menggunakan hyperparameter terbaik {’C’: 2,’epsilon’: 0,2, ’gamma’: ’scale’, ’kernel’: ’rbf’, ’tol’: 0,005} untuk dataset pria, dan {’C’: 3, ’epsilon’: 0,2, ’gamma’: ’scale’, ’kernel’: ’rbf’, ’tol’: 0,005} untuk dataset wanita. Metode RF memiliki kinerja yang cukup baik, dengan nilai RMSE = 7,532; R2 = 0,403; Adjusted R2 = 0,351; running time = 0,154 untuk pria dan RMSE = 6,889; R2 = 0,340; Adjusted R2 = 0,264; running time = 0,179 untuk wanita. Selain itu, SVR juga memiliki performa yang cenderung sama namun sedikit lebih buruk, dengan nilai RMSE = 7,692; R2 = 0,376; Adjusted R2 = 0,321; running time = 0,035 untuk pria dan RMSE = 6,905; R2 = 0,337; Adjusted R2 = 0,306; running time = 0,080 untuk wanita. Berdasarkan analisis kinerja model yang dilakukan pada penelitian ini model yang

dibangun dengan Metode Random Forest Regression lebih unggul dalam memprediksi
usia biologis dibandingkan dengan Metode Support Vector Regression.
.....Aging is one of the main risk factors for disease and death. The aging rate of individ- uals of the same chronological age has been shown to vary. So therefore, a need arises for a more accurate, robust, and reliable aging measurement tool than chronological age, namely biological age. In this research, the author build a model using the Random For- est Regression (RF) Method and the Support Vector Regression (SVR) Method to predict biological age from patient clinical data, assess and evaluate the performance results, and compare the performance of the two models. Regarding the method used, the Random Forest Regression Method is a method that applies the Ensemble Learning Technique by combining several decision trees to produce predictions. Meanwhile, the Support Vector Regression Method is a method that works by building a hyperplane or collection of hy- perplane in high-dimensional space which can be used for linear or nonlinear regression. The dataset used is medical data originating from the Ministry of Health of the Republic of Indonesia. On the dataset, data preprocessing is carried out, namely the data is processed in the aspects of missing values handling, encoding, and outliers detection and outliers handling. Then, feature selection is carried out using Spearman's Rank Correlation Co- efficient. After that, machine learning model using RF Method and machine learning model using SVR Method were created separately for each gender. Finally, the model performance is evaluated and its performance compared using evaluation metrics, namely Root Mean Square Error (RMSE), Coefficient of Determination (R2), and Adjusted R2, as well as running time. The RF Method used best hyperparameters {'max depth': 15, 'n estimators': 1150} for the male dataset, and {'max depth': 15, 'n estimators': 1250 } for the female dataset. Meanwhile, the SVR Method used best hyperparameters {'C': 2, 'epsilon': 0.2, 'gamma': 'scale', 'kernel': 'rbf', 'toll': 0.005} for the male dataset, and {'C': 3, 'epsilon': 0, 2, 'gamma': 'scale', 'kernel': 'rbf', 'toll': 0.005} for female dataset. The result is that the model built using the RF Method has quite good performance, with an RMSE value of = 7.532; R2 = 0.403; Adjusted R2 = 0.351; running time = 0.154 for men and RMSE = 6.889; R2 = 0.340; Adjusted R2 = 0.264; running time = 0.179 for women. Apart from that, SVR also has performance that tends to be the same but slightly worse, with an RMSE value of = 7,692; R2 = 0.376; Adjusted R2 = 0.321; running time = 0.035 for men and RMSE = 6.905; R2 = 0.337; Adjusted R2 = 0.306; running time = 0.080 for women. Based on the model performance analysis carried out in this research, the model built using the Random Forest Regression Method is superior in predicting biological age compared to the Support Vector Regression Method.