INTISARI

PREDIKSI BIAYA ASURANSI DENGAN MENGGUNAKAN *MACHINE LEARNING* REGRESI LINEAR BERGANDA

Oleh

LYRA AYUKUSUMANINGTYAS

20/462284/PA/20256

Skripsi ini bertujuan untuk memprediksi biaya asuransi yang dipengaruhi oleh faktor-faktor seperti umur, jenis kelamin, perokok, BMI, jumlah anak, dan wilayah. Untuk mengidentifikasi variabel-variabel yang memiliki pengaruh signifikan terhadap biaya asuransi digunakan analisis korelasi. Analisis korelasi berguna untuk menentukan hubungan antara variabel independen, seperti umur, jenis kelamin, perokok, BMI, jumlah anak, dan wilayah dengan variabel dependennya yaitu biaya. Dengan mengetahui besarnya korelasi, dapat dinilai seberapa erat hubungan variabel-variabel independen tersebut dengan variabel dependen (biaya asuransi). Hasil analisis korelasi ini akan membantu dalam memilih variabel yang paling relevan untuk dimasukkan ke dalam model regresi. Berdasarkan hasil analisis, diperoleh tiga variabel yang memiliki pengaruh signifikan terhadap biaya, yaitu perokok, umur, dan BMI. Model regresi linear berganda akan dibangun dengan menggunakan ketiga variabel independen tersebut. Data yang diolah dalam penelitian ini diperoleh dari *www.kaggle.com* yang terdiri dari 1338 baris dan 7 kolom. Data dibagi menjadi 70% data pelatihan dan 30% data pengujian. Proses dalam menentukan model prediksi dilakukan menggunakan *Google Colaboratory* dengan *library* seperti NumPy, Pandas, Matplotlib, Seaborn, dan scikit-learn. Proses prediksi biaya asuransi dilakukan melalui beberapa tahapan, yaitu menyiapkan *library* yang akan digunakan, menganalisis data, melakukan data *pre-processing*, memilih data, dan melakukan proses evaluasi. Proses ini menghasilkan model regresi linear berganda dengan masing-masing koefisien pada variabel perokok, umur, and BMI. Model dievaluasi dengan menggunakan metrik seperti nilai koefisien determinasi (R^2), MSE, RMSE, MAE, dan MAPE. Hasil dari evaluasi model menunjukkan bahwa model regresi linear berganda yang diperoleh baik untuk memprediksi biaya asuransi. Evaluasi akhir yang dilakukan digunakan untuk memberikan gambaran tentang perbandingan antara nilai aktual y dan nilai prediksi $\hat{y}$ dalam bentuk tabel dan grafik.

ABSTRACT

INSURANCE COST PREDICTION USING MACHINE LEARNING MULTIPLE LINEAR REGRESSION

By

LYRA AYUKUSUMANINGTYAS

20/462284/PA/20256

The undergraduate thesis aims to predict insurance costs influenced by factors such as age, gender, smoking status, BMI, number of children, and region. To identify the variables that have a significant impact on insurance costs, correlation analysis is used. This analysis helps to determine the relationships between independent variables, such as age, gender, smoking status, BMI, number of children, and region and the dependent variable, which is the insurance cost. By assesing the strength of these correlations, we can evaluate how closely each independent variable is related to dependent variable (insurance costs). The results of this correlation analysis will aid in selecting the most relevant variables to include in the regression model. Based on the analysis, three variables were found to have a significant impact on insurance costs, smoking status, age, and BMI. A multiple linear regression model will be developed using these three independent variables. The processed data was obtained from www.kaggle.com, consisting of 1338 rows and 7 columns. The data was split into 70% training data and 30% testing data. The process of developing the predictive model is carried out using Google Colaboratory, with libraries such as NumPy, Pandas, Matplotlib, Seaborn, and scikit-learn. The insurance cost prediction process was carried out through several stages, namely preparing the necessary libraries, analyzing the data, performing data pre-processing, selecting data, and conducting the evaluation process. This process resulted in a multiple linear regression model with coefficients corresponding to the variables of smoking status, age, and BMI. The model evaluated using metrics such as the coefficient of determination (R^2), MSE, RMSE, MAE, and MAPE. The results of the model evaluation indicate that the obtained multiple linear regression model is effective for predicting insurance costs. The final evaluation is conducted to provide an overview of the comparison between the actual values y and the predicted values $\hat{y}$ in the form of tables and graphs.