Project Title : Predicting Medical Insurance Costs Using Machine Learning

Title : A Computational Intelligence Approach for Predicting Medical Insurance Cost

Authors : Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeeb A. A. Mosleh, Syed Sajid Ullah

DOI : https://doi.org/10.1155/2021/1162553

Affiliations : Capital University of Science and Technology, Islamabad, Pakistan, Universiti Brunei Darussalam, Brunei Darussalam, King Saud University, Riyadh, Saudi Arabia, Taiz University, Yemen, Villanova University, PA, USA

Published 28 December 2021

Abstract :

This study presents a computational intelligence approach leveraging machine learning (ML) techniques to predict healthcare insurance costs. By utilizing various regression models on a dataset from Kaggle, the research demonstrates the effectiveness of these models in forecasting medical insurance expenses, with Stochastic Gradient Boosting (SGB) showing the highest accuracy.

Methodology:

Dataset : The medical insurance cost dataset from Kaggle, containing attributes such as age, sex, BMI, children, smoker status, region, and charges.

Data Preprocessing : Conversion of categorical data to numerical values, feature engineering, and splitting the dataset into training (70%) and testing (30%) sets.

ML Models Used :

- Linear Regression (LR)
- Support Vector Regression (SVR)
- Ridge Regressor (RR)
- Stochastic Gradient Boosting (SGB)
- XGBoost (XGB)
- Decision Tree (CART)
- Random Forest Regressor (RFR)
- Multiple Linear Regression (MLR)
- k-Nearest Neighbors (kNN)

Evaluation Metrics : Models were evaluated using cross-validation scores and Root Mean Squared Error (RMSE).

Results and Analysis :

Performance Comparison : SGB model achieved the highest accuracy with a cross-validation score of 0.858 and an RMSE of 0.340, translating to 86% accuracy.

Feature Importance : Attributes like age, BMI, and smoker status were found to be the most significant predictors of medical insurance costs.

Conclusion :

This study demonstrates that machine learning models, particularly ensemble methods like Stochastic Gradient Boosting, can effectively predict healthcare insurance costs. These findings can guide the development of robust predictive systems for the healthcare industry, aiding in better resource allocation and planning.