Medical Insurance Cost Prediction

End-to-End Machine Learning Pipeline

2026-01-01

Problem Overview

This report presents the results of an end-to-end machine learning pipeline developed to predict medical insurance costs based on personal and geographic attributes. The solution includes data ingestion, model training, evaluation, scoring, and automated reporting.

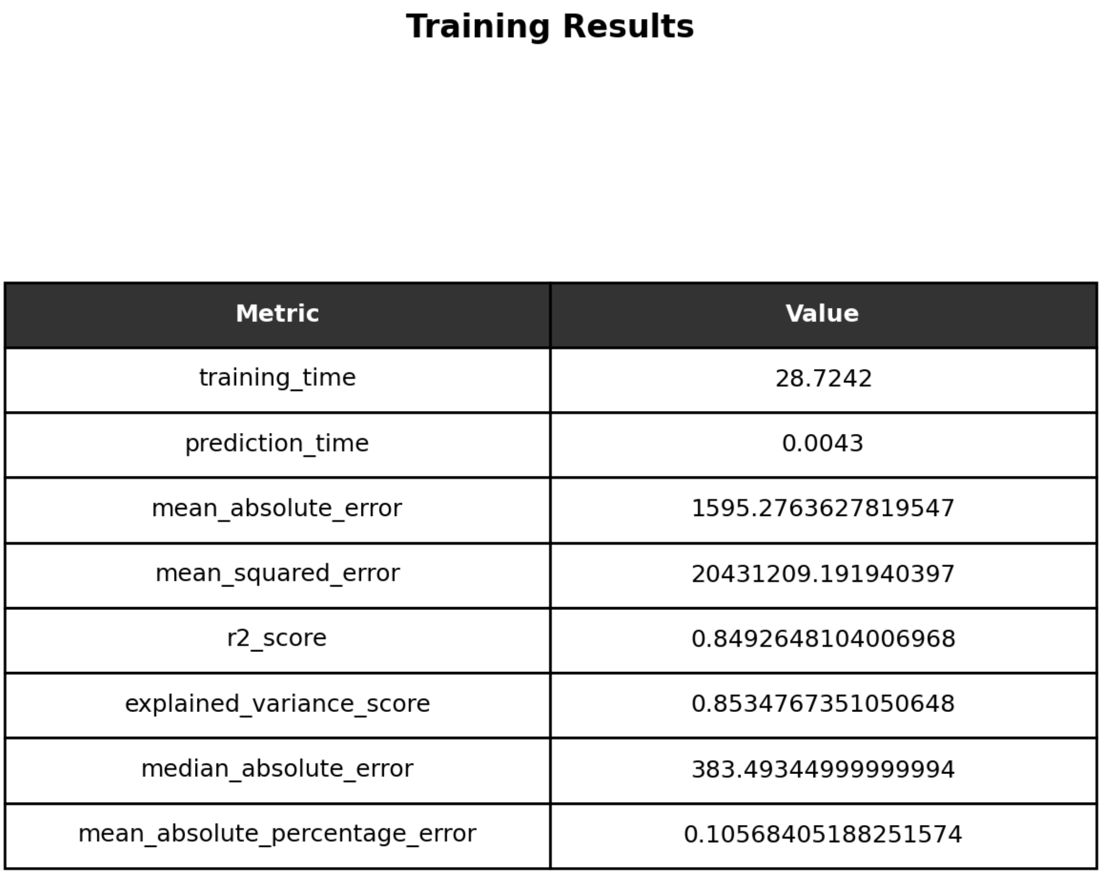
Dataset Overview

- Number of samples: 1338
- Target variable: Medical insurance charges
- Features include demographic, health, and regional attributes

Model Information

- Model type: DecisionTreeRegressor
- Random state: 42
- Training strategy: supervised regression

Training Metrics

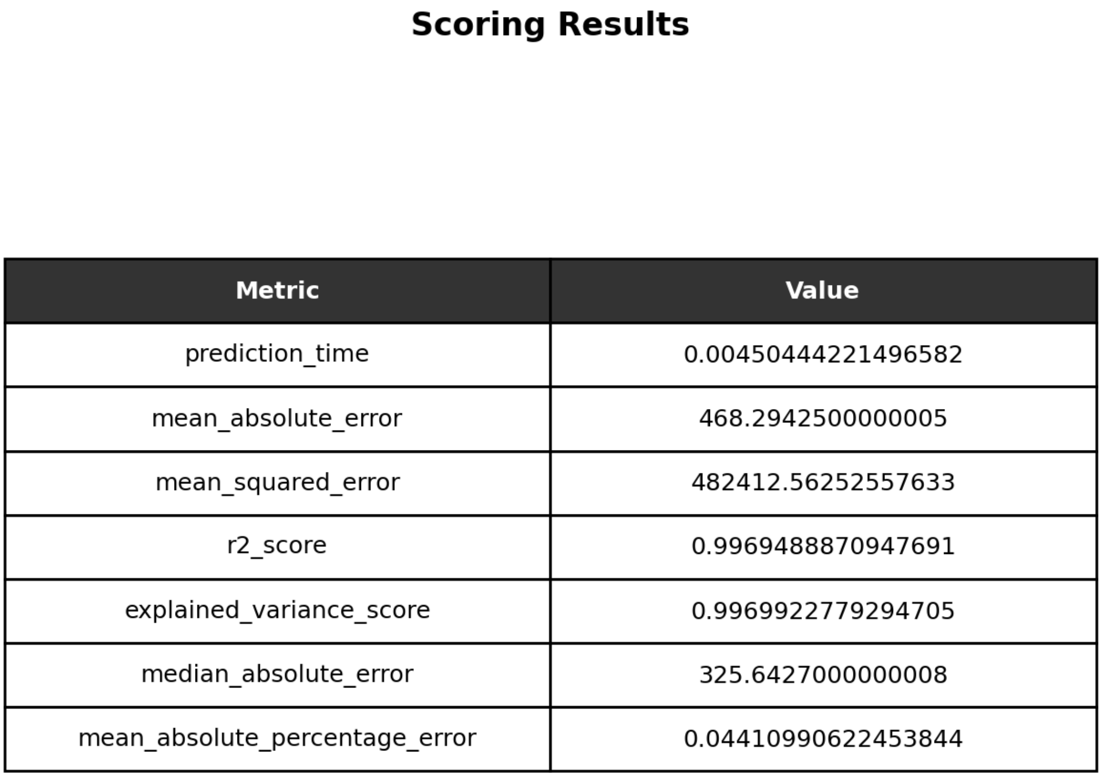


Training Results

Metric	Value
training_time	28.7242
prediction_time	0.0043
mean_absolute_error	1595.2763627819547
mean_squared_error	20431209.191940397
r2_score	0.8492648104006968
explained_variance_score	0.8534767351050648
median_absolute_error	383.49344999999994
mean_absolute_percentage_error	0.10568405188251574

Figure 1: Training metrics performance

Validation Metrics

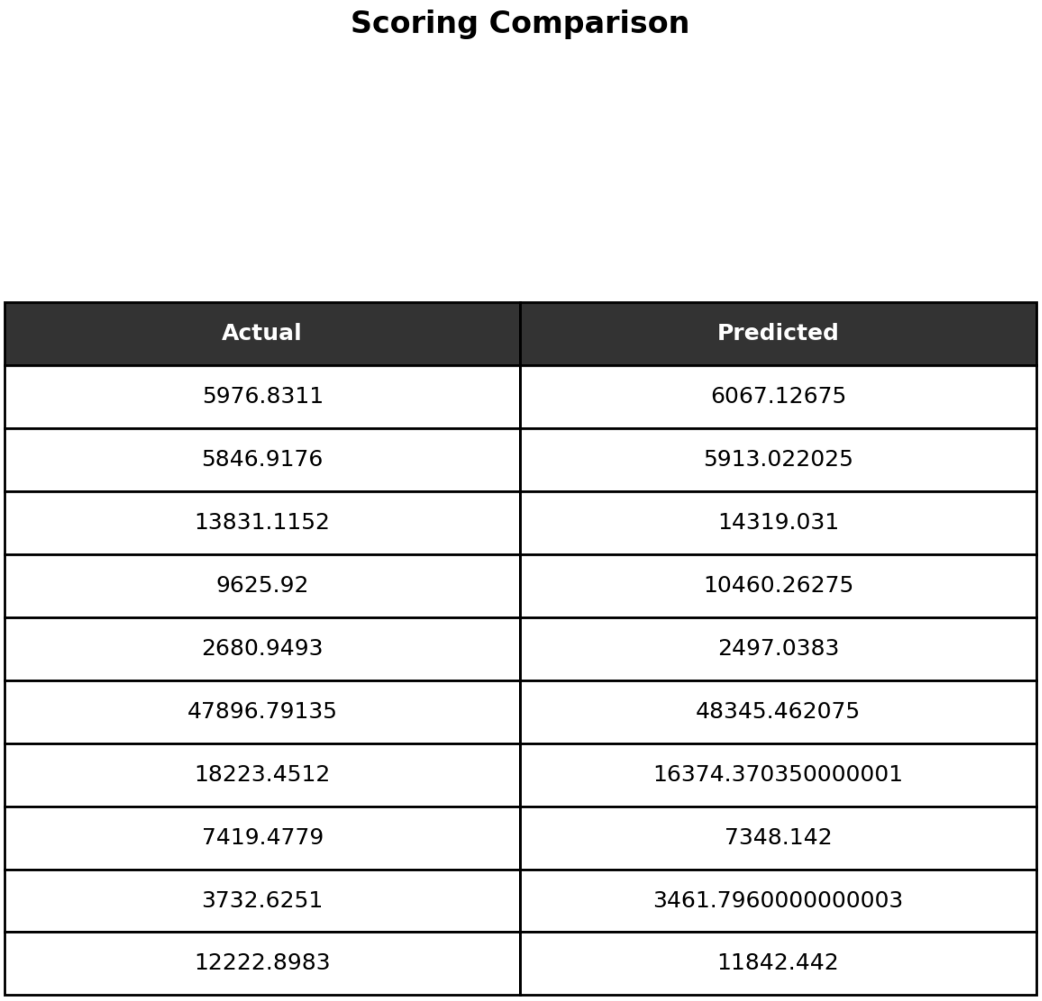


Scoring Results

Metric	Value
prediction_time	0.00450444221496582
mean_absolute_error	468.2942500000005
mean_squared_error	482412.56252557633
r2_score	0.9969488870947691
explained_variance_score	0.9969922779294705
median_absolute_error	325.6427000000008
mean_absolute_percentage_error	0.04410990622453844

Figure 2: Validation metrics performance

Predictions vs Actual Values



Scoring Comparison

Actual	Predicted
5976.8311	6067.12675
5846.9176	5913.022025
13831.1152	14319.031
9625.92	10460.26275
2680.9493	2497.0383
47896.79135	48345.462075
18223.4512	16374.370350000001
7419.4779	7348.142
3732.6251	3461.7960000000003
12222.8983	11842.442

Figure 3: Comparison between real and predicted insurance charges

Final Evaluation

The trained model was evaluated on a hold-out dataset generated through random sampling. The results demonstrate the model's ability to capture the underlying patterns in medical insurance costs while maintaining generalization performance.