Medical Insurance Cost Prediction

End-to-End Machine Learning Pipeline

2026-01-03

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: prueba$_g if$
- Random state: 42
- Training strategy: supervised regression

Training Metrics

Training Results for prueba_gif

Metric	Value
training_time	4.4022
prediction_time	0.0062
mean_absolute_error	2975.264169021501
mean_squared_error	30001073.94219962
r2_score	0.8081158615936221
explained_variance_score	0.8086121026095596
median_absolute_error	1420.6563391904122
mean_absolute_percentage_error	0.275654085533741

Figure 1: Training metrics performance

Validation Metrics

Scoring Results for prueba_gif

Metric	Value
prediction_time	0.0057909488677978516
mean_absolute_error	1823.3871220003216
mean_squared_error	5014538.828288441
r2_score	0.9528612620796859
explained_variance_score	0.9693371215732991
median_absolute_error	1666.3363611455115
mean_absolute_percentage_error	0.28204335984163864

Figure 2: Validation metrics performance

Predictions vs Actual Values

Scoring Comparison for prueba_gif

Actual	Predicted
9095.06825	9899.371109062502
5272.1758	6904.984556780823
29330.98315	26911.36211333333
9301.89355	9899.371109062502
33750.2918	35553.87148608696
4536.259	9158.708791578947
2117.33885	2039.4538827956985
14210.53595	15350.43995471698
3732.6251	7168.603693333332
10264.4421	11964.3060655102

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.