Medical Insurance Cost Prediction

A Comprehensive Analysis with Machine Learning

Hasti Aksoy

September 28, 2025

Abstract

Accurate prediction of medical insurance charges is an essential problem for both insurance providers and individuals. By modeling demographic and lifestyle attributes, we can anticipate healthcare costs and assess risk. In this project, we analyze the *Medical Cost Personal Dataset* (N=1,338; 7 attributes) using a structured pipeline: data preprocessing, exploratory data analysis (EDA), model training, evaluation, and model selection. We compare multiple models—linear regression, decision tree, random forest, and gradient boosting (XGBoost)—and find that the Random Forest Regressor provides the best balance of accuracy and efficiency, achieving RMSE = 4228.69 and $R^2 = 0.847$. This report details the methodology, experiments, results, and recommendations for future work.

Contents

1 Introduction

Health insurance plays a central role in managing healthcare expenses, yet predicting individual costs remains a challenge. Costs are influenced by age, body composition, lifestyle habits, and smoking status. Insurance providers rely on accurate cost prediction to set premiums fairly, while individuals benefit from understanding how their demographics and choices affect financial outcomes.

The dataset used in this project, commonly called the *Medical Cost Personal Dataset*, consists of 1,338 records. Each record includes:

- **Age:** integer between 18 and 64,
- **Sex:** male or female,
- **BMI:** body mass index, a continuous measure of body fat,
- **Children:** number of dependents,
- **Smoker:** binary indicator (yes/no),
- **Region:** four geographic categories,
- **Charges:** medical insurance cost (target variable).

The project's goal is to preprocess and analyze the dataset, identify important predictors, and build machine learning models that provide reliable estimates of medical costs.

2 Data Preprocessing

High-quality data is crucial for effective modeling. To ensure consistency and reproducibility, a dedicated `DataCleaning` class was implemented. The following steps were performed:

1. **Text canonicalization:** all categorical strings were lowercased and stripped of whitespace; region names were standardized (e.g., "south west" → "southwest").
2. **Duplicate removal:** one duplicate row was detected and removed, resulting in $N = 1337$ unique records.
3. **Schema validation:** numeric columns were checked against expected ranges (e.g., age between 0–120, positive BMI, non-negative charges). Invalid rows were dropped.
4. **Missing values:** categorical columns were filled with their mode, and numeric columns with the median. Group-wise medians (based on `sex`, `smoker`, `region`) were used where appropriate.
5. **Outlier handling:** the interquartile range (IQR) method was applied to `bmi` and `charges`. Outliers were winsorized rather than dropped, preserving dataset size.
6. **Encoding:** `sex` and `smoker` were label-encoded using fixed mappings (female=0, male=1; no=0, yes=1). The `region` column was one-hot encoded with one category dropped to avoid multicollinearity.

After preprocessing, the cleaned dataset was saved as `insurance_cleaned.csv`.

3 Exploratory Data Analysis (EDA)

Exploratory analysis revealed the following key insights:

Distributions

Age was broadly distributed across adults; BMI approximated a normal distribution with slight skew; charges were highly right-skewed with a heavy tail of very high-cost individuals.

Gender

Males had slightly higher average charges than females, but gender alone was not strongly predictive.

Smoking status

Smoking emerged as the most significant factor. Smokers had dramatically higher charges, with average values multiple times greater than non-smokers. Male smokers had the highest costs overall.

Age $\times$ Smoking

Among non-smokers, charges increased gradually with age. Among smokers, charges increased sharply with age, suggesting a compounded effect of aging and smoking.

BMI $\times$ Smoking

Obese smokers exhibited the highest average charges of any group, reinforcing the interaction between lifestyle and health outcomes.

Children

The number of dependents did not show a consistent linear relationship with charges, though variance was greater among families with more children.

Region

Slight differences were observed, with the Southeast showing higher average charges, but the effect was small compared to smoking or BMI.

Correlation analysis

Correlation coefficients confirmed:

- Smoker ($r \approx 0.79$),
- Age ($r \approx 0.31$),
- BMI ($r \approx 0.16$).

Other features, including gender, children, and region, had weak correlations with charges.

Figures

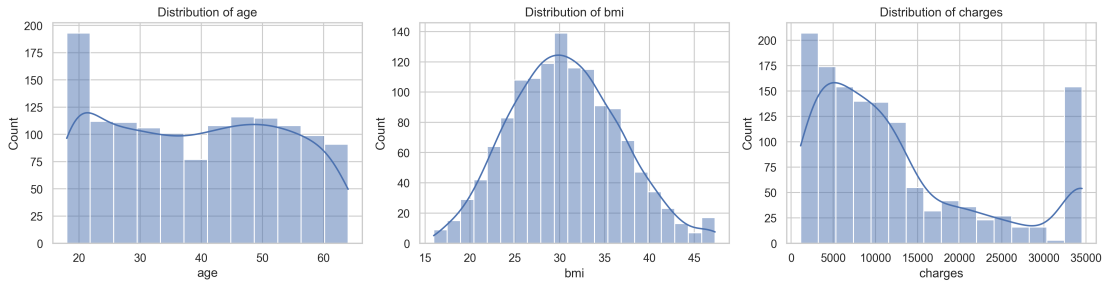


Figure 1: Distributions of age, BMI, and charges.

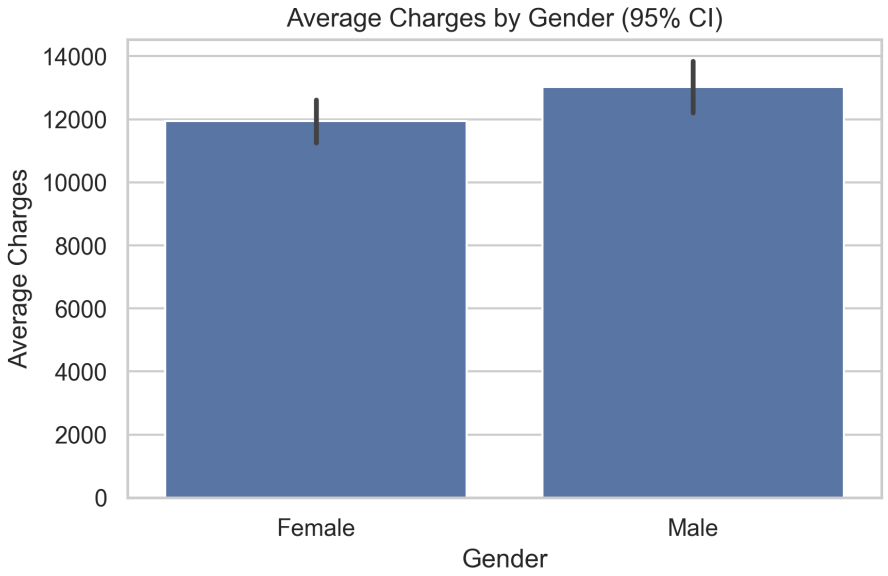
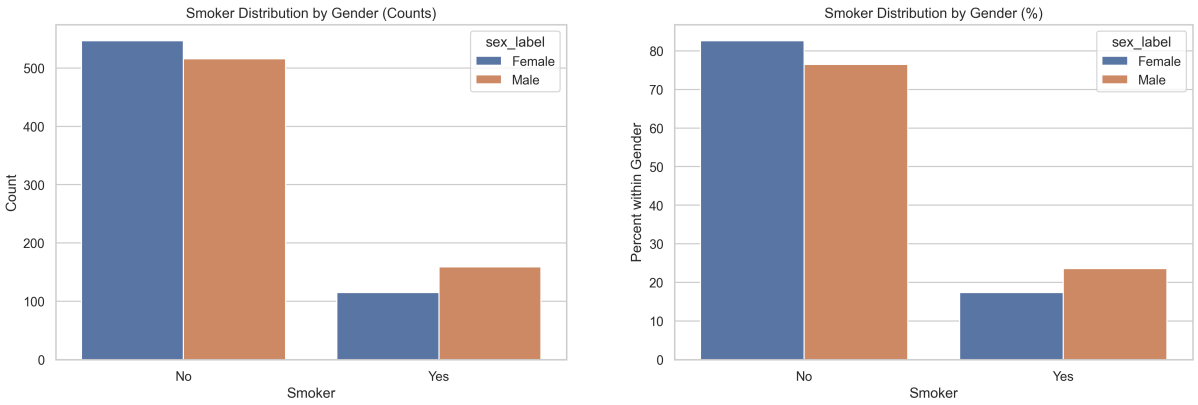


Figure 2: Average charges by gender.



(a) Counts by gender.

(b) Percent by gender.

Figure 3: Smoking distribution by gender.

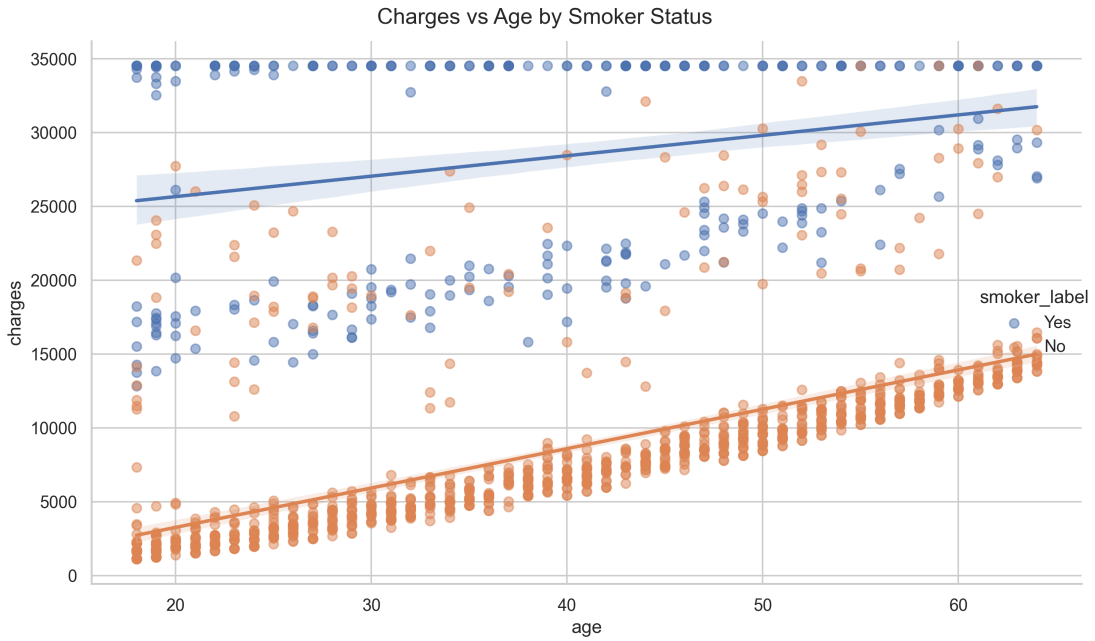


Figure 4: Age vs charges by smoking status.

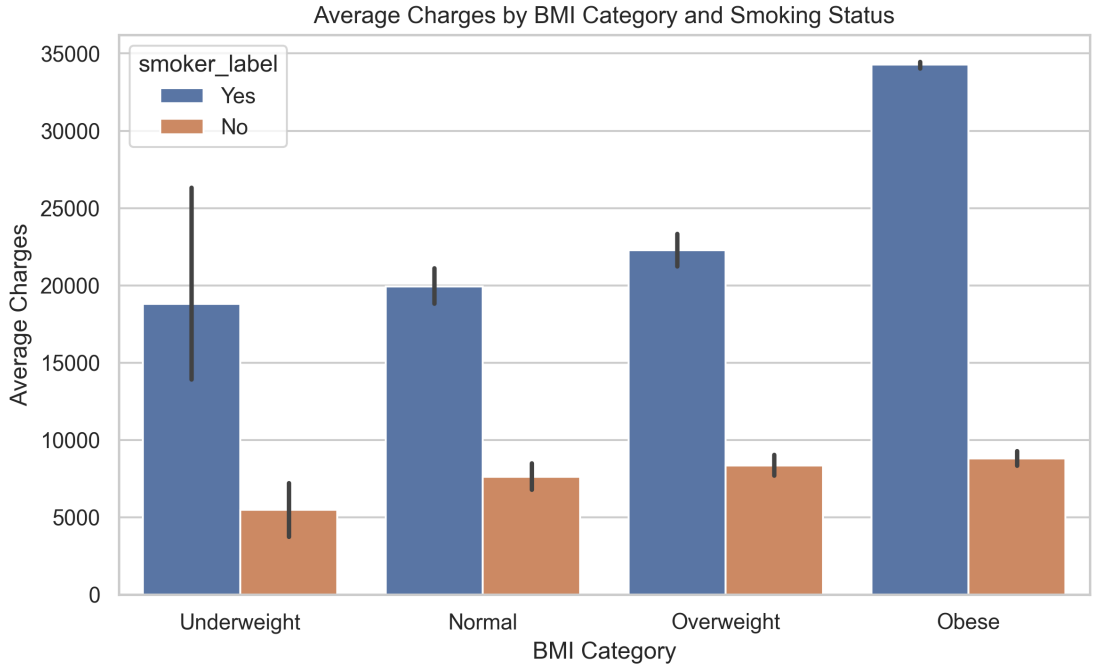


Figure 5: Average charges by BMI category and smoker.

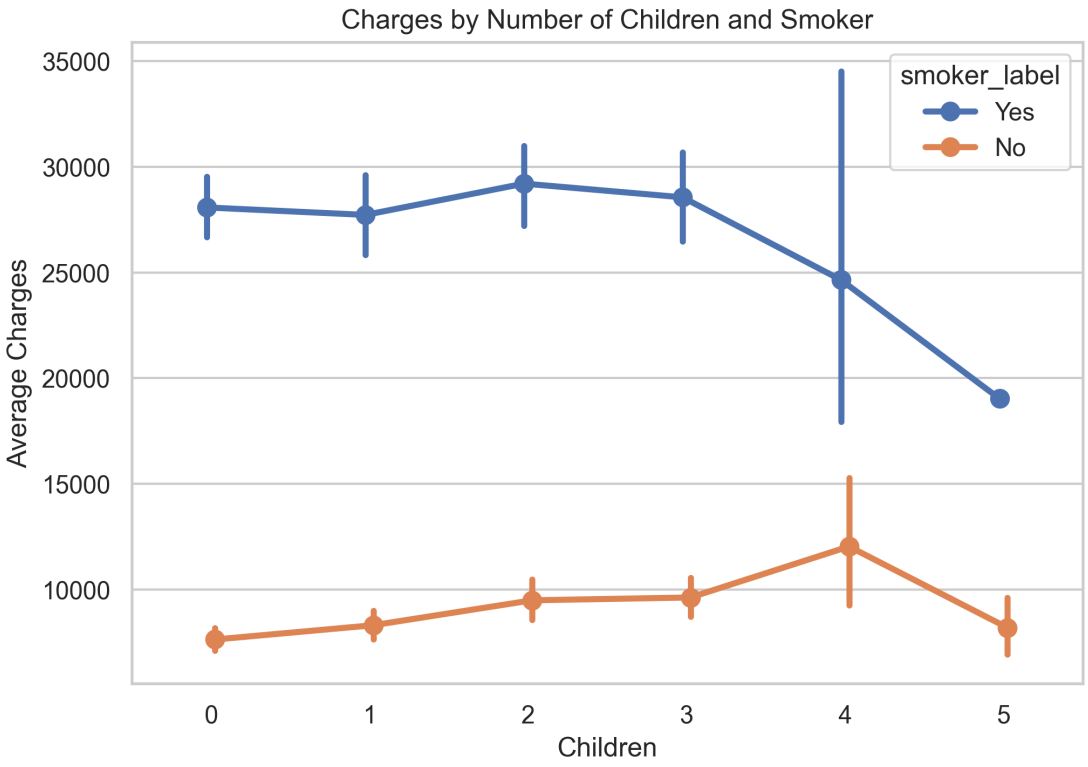


Figure 6: Average charges by number of children and smoker status.

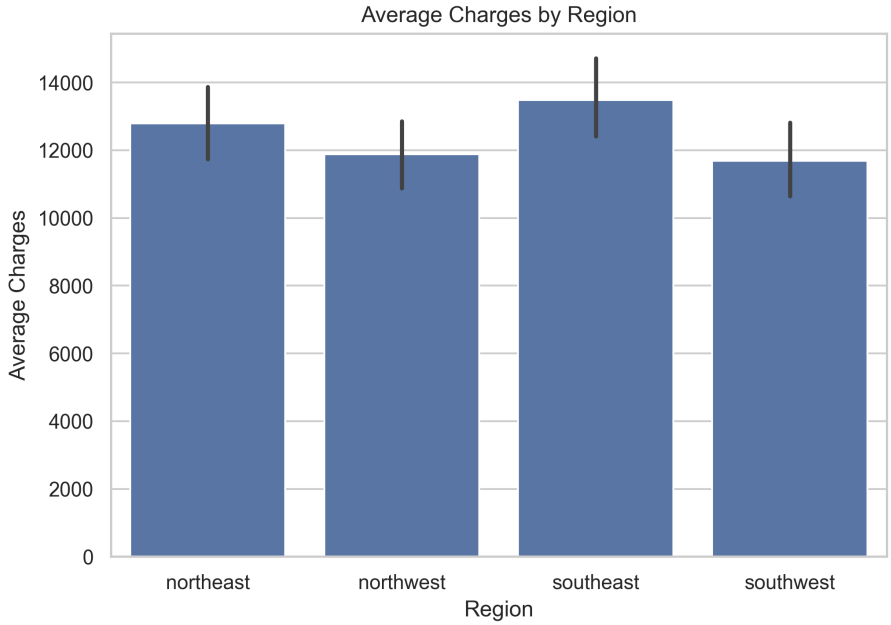
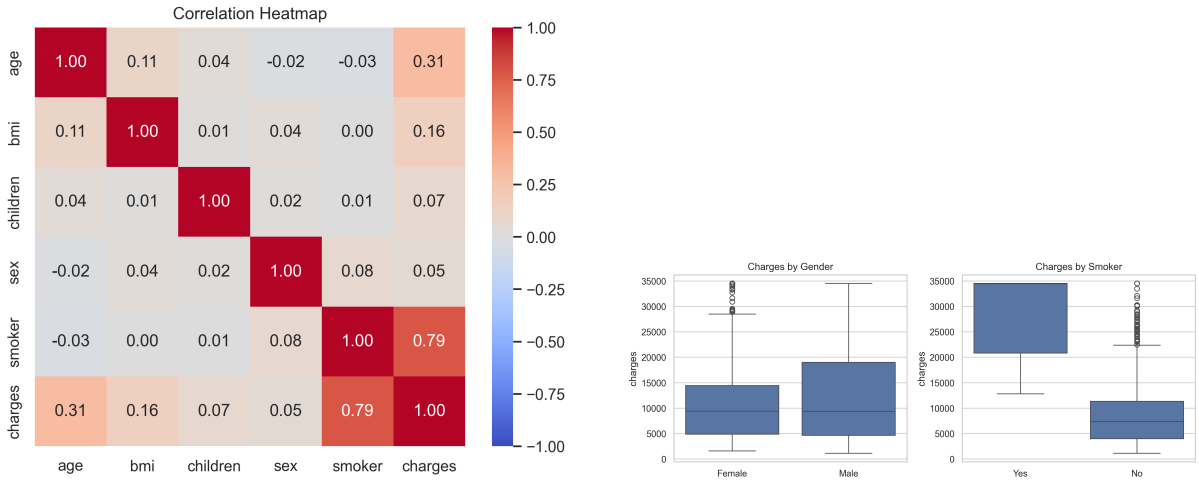


Figure 7: Average charges by region.



(a) Correlation heatmap.

(b) Boxplots by gender and smoker.

Figure 8: Correlation structure and target distribution across key groups.

4 Methods

The prediction task was framed as supervised regression.

4.1 Baseline: Linear Regression

A standard linear regression model was used as the baseline. Preprocessing (scaling via `StandardScaler`) was encapsulated in a pipeline.

4.2 Tree-Based Models

- **Decision Tree Regressor:** maximum depth of 6 to prevent overfitting.
- **Random Forest Regressor:** ensemble of 300 trees, providing variance reduction and robustness.

4.3 Gradient-Boosted Trees

XGBoost was configured with 500 estimators, learning rate 0.05, maximum depth of 5, subsample 0.8, and `colsample_bytree` 0.8.

4.4 Training Protocol

- Dataset split: 80/20 train–test.
- Baseline model additionally evaluated with 5-fold cross-validation on training data.
- Evaluation metrics: RMSE (primary), RMSLE (sensitive to skew), MAE, and R^2.
- Wall-clock training and prediction times were also recorded.

5 Results

5.1 Performance Metrics

Model	RMSE	RMSLE	MAE	R^2	Fit (s)	Predict (s)	Total (s)
LinearRegression	4572.25	0.43	3151.60	0.82	0.02	0.00	0.02
DecisionTree	4271.39	0.47	2267.77	0.84	0.02	0.01	0.03
RandomForest	**4228.69**	0.47	**2242.10**	**0.8470**	0.71	0.08	0.79
XGBoost	4365.77	0.68	2582.44	0.84	2.66	0.01	2.67

Table 1: Holdout results on the 20% test split (80/20 train–test).

5.2 Interpretation

- **Random Forest** achieved the lowest RMSE (4228.69) and highest R^2 (0.847), making it the best-performing model.
- **Linear regression** underperformed, indicating that linear relationships alone are insufficient.
- **Decision Tree** performed better than linear regression but was less stable.
- **XGBoost** did not outperform Random Forest under the tested configuration and required more training time.
- **RMSLE** was lowest for linear regression, suggesting Random Forest tended to over-predict lower-cost cases.

6 Model Manifest

The selected best model (Random Forest) was saved as `models/best_model.joblib`. Its manifest is:


```
{
  "name": "RandomForest",
  "metrics": {
    "rmse": 4228.69,
    "rmsle": 0.4734,
    "mae": 2242.10,
    "r2": 0.8470,
    "fit_seconds": 0.7095,
    "predict_seconds": 0.0766,
    "total_seconds": 0.7861
  }
}
```

7 Discussion

Tree-based models clearly outperform linear regression, consistent with EDA findings of non-linear feature interactions (e.g., smoker × BMI). Random Forest was the most balanced model, providing strong accuracy without excessive runtime.

XGBoost, although theoretically powerful, did not surpass Random Forest in this case, likely due to limited hyperparameter tuning. The slight increase in RMSLE for ensemble models suggests systematic overestimation for low charges, highlighting potential areas for calibration.

8 Conclusion

- Random Forest achieved the best results (RMSE = 4228.69, MAE = 2242.10, R^2 = 0.847).
- Smoking, age, and BMI were the strongest predictors of insurance charges.
- The cleaned dataset and models were successfully encapsulated in pipelines for reproducibility.

9 Future Work

To extend and improve this work:

1. Apply log-transformation of charges to stabilize skew and improve RMSLE.
2. Perform hyperparameter tuning for Random Forest and XGBoost.
3. Use SHAP or permutation importance for interpretability.
4. Evaluate with full k-fold cross-validation to ensure robustness.
5. Explore residual analysis and calibration techniques to address systematic biases.