# Health Insurance Premium Prediction

## Problem Statement

The task at hand is to predict health insurance premiums (`charges`) for individuals based on their demographic, medical, and lifestyle attributes. The dataset includes features such as age, gender, weight, BMI, hereditary diseases, number of dependents, smoking habits, blood pressure, diabetes, regular exercise, and job title. The primary objective is to develop a model that accurately predicts the insurance charges, helping to assist insurance companies in estimating premiums based on individual profiles. By understanding the factors that influence insurance premiums, the model can also provide insights into how different variables affect premium costs.

## Methodology

1. \*\*Data Preprocessing:\*\*  
 - Missing values were handled by filling numerical features with the mean and categorical features with the mode.  
 - Categorical variables were encoded using one-hot encoding.  
 - Features were scaled to ensure uniformity in feature ranges.  
  
2. \*\*Feature and Target Separation:\*\*  
 - Independent features included demographic, medical, and lifestyle attributes.  
 - The target variable (`charges`) represented the health insurance premium.  
  
3. \*\*Train-Test Split:\*\*  
 - Data was divided into training, validation, and test sets to ensure robust evaluation.  
  
4. \*\*Model Development:\*\*  
 - A Random Forest Regressor was used for its ability to handle non-linear relationships and robustness to overfitting.  
 - Grid Search was employed for hyperparameter tuning.  
  
5. \*\*Model Evaluation:\*\*  
 - Metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² Score were used.  
 - Residual analysis and visualizations were performed to validate model accuracy.  
  
6. \*\*Visualization:\*\*  
 - Plots included Actual vs Predicted values, residual distributions, and feature importances.  
  
7. \*\*Model Saving:\*\*  
 - The final model was saved using `joblib` for future use.

## Results

1. \*\*Hyperparameter Tuning Results:\*\*  
 - The best hyperparameters were identified through Grid Search, optimizing model performance.  
  
2. \*\*Validation and Test Set Performance:\*\*  
 - Validation and test metrics showed the model generalizes well to unseen data:  
 - Low Mean Absolute Error (MAE) and Mean Squared Error (MSE).  
 - High R² Score, indicating the model explains a significant portion of variance in premiums.  
  
3. \*\*Feature Importance:\*\*  
 - Key predictors of insurance premiums included `age`, `smoker`, `bmi`, and `weight`.  
  
4. \*\*Model Performance Visualizations:\*\*  
 - Actual vs Predicted values indicated accurate predictions.  
 - Residual plots confirmed no significant bias in predictions.  
 - Feature importance analysis aligned with expectations, showing lifestyle and demographic factors heavily influence premiums.  
  
5. \*\*Model Saving:\*\*  
 - The trained model was saved successfully and is ready for deployment.

## Conclusion

The Random Forest Regressor model effectively predicts health insurance premiums based on demographic, medical, and lifestyle attributes. With accurate predictions and robust generalization, this model provides valuable insights into the key factors influencing premiums and can assist insurance companies in estimating costs for individual profiles. The model can be further optimized or extended with additional data for improved performance.