📝 Health Insurance Cost Prediction — Project Report

1. Introduction

Health insurance plays a pivotal role in safeguarding individuals from high medical expenses. Accurately predicting insurance costs can help insurance providers design better policies, offer dynamic pricing, and minimize losses. This project focuses on predicting medical charges using demographic and health-related features through machine learning models.

2. Business Problem

Insurance companies face challenges in accurately estimating medical costs due to variable lifestyle, demographic, and behavioral factors. Underestimation leads to loss, while overestimation may drive away customers.

Key Questions:

- Can we identify which factors influence insurance charges?
- Can we predict charges for a new applicant?
- How can the company use this to optimize revenue?

3. Dataset Overview

Source: `insurance.csv`
Records: 1338
Features:

- age: Age of the insured person
- sex: Gender of the person

- `bmi`: Body mass index (health risk indicator)
- `children`: Number of children covered
- `smoker`: Smoking status
- `region`: Residential area (southeast, southwest, etc.)
- `charges`: Medical insurance premium (target variable)

4. Exploratory Data Analysis (EDA)

- **Age vs Charges**: Costs increase steadily with age, especially after 40.
- **Smoking**: Smokers are charged significantly higher premiums — nearly 2–3x more.
- **BMI**: Higher BMI (esp. >30) correlates with higher insurance cost.
- **Gender and Region**: Minimal impact on charges.
- **Children**: Slight increase with more children, but not significant.

Outliers: Detected especially in charges (due to high-risk smokers).

5. Data Preprocessing

- Categorical features encoded:
  - Label Encoding (`sex`, `smoker`)
  - One-Hot Encoding (`region`)
- Feature scaling done using **StandardScaler**
- No missing values observed.

6. Modeling

Models tried:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regressor ✅ **(Best Performance)**

Model Performance:

- **Random Forest R^2 Score:** ~0.87
- **RMSE (Root Mean Squared Error):** ~4300

7. Insights

- **Smoking** is the most influential feature, dramatically increasing premiums.
- **Age and BMI** also have strong effects on cost.
- Gender and region have **negligible influence**.
- Maintaining a healthy BMI and quitting smoking could substantially reduce costs.

8. Recommendations

1. **Introduce dynamic pricing**: Based on health status (BMI, smoking).
2. **Customer education**: Promote healthy lifestyle campaigns.
3. **Wellness incentives**: Discounts for non-smokers or healthy BMI individuals.
4. **Predictive tools**: Deploy the model on the website for real-time premium quotes.

9. Possible Enhancements

- Add more granular data (e.g., exercise, diet, past illnesses).
- Use advanced models (XGBoost, CatBoost).
- Feature selection automation (e.g., Recursive Feature Elimination).
- Explainability tools like SHAP/LIME to interpret model behavior.
- Outlier treatment with IQR or robust scaling.
- Web deployment via Flask or Streamlit.

10. Conclusion

This project successfully demonstrates the feasibility of using machine learning to estimate health insurance costs. With a well-performing Random Forest Regressor, we can make reliable predictions and offer valuable insights into risk-based pricing strategies.