Insurance Charge Prediction

# 1. Executive Summary

This project focuses on predicting medical insurance charges for individuals based on demographic and health-related factors such as age, BMI, smoking habits, gender, and region. By applying Exploratory Data Analysis (EDA), feature engineering, and machine learning models, the aim is to estimate insurance costs with accuracy. The results help insurance providers in risk assessment and premium calculation.

# 2. Problem Statement

Insurance companies face challenges in fairly pricing premiums for individuals while maintaining profitability. An accurate predictive model can help determine expected insurance charges by analyzing customer profiles.

# 3. Objectives

- Perform data cleaning and preprocessing  
- Conduct EDA to extract insights  
- Engineer features to improve prediction  
- Train and compare ML models for insurance charge prediction  
- Evaluate performance using error metrics

# 4. Dataset Description

Source: [Opensource Website]  
Features:  
- Age (numeric)  
- Sex (categorical)  
- BMI (numeric)  
- Children (numeric)  
- Smoker (categorical)  
- Region (categorical)  
- Charges (target variable)

# 5. Exploratory Data Analysis (EDA)

- Distribution plots for Age, BMI, and Charges  
- Boxplots for Charges vs. categorical features (Smoker, Region, etc.)  
- Correlation heatmap to identify strongest predictors (e.g., BMI, Smoking)  
- Insights: Smoking and higher BMI significantly increase charges

# 6. Data Preprocessing & Feature Engineering

- Handled missing values (if any)  
- Encoded categorical variables   
- Scaled numerical features where required  
- Checked multicollinearity between predictors

# 7. Model Building

Tested models:  
- Linear Regression – baseline model

# 8. Model Evaluation

Metrics used:   
- Adjusted R-squared   
- RMSE (Root Mean Squared Error)  
- R² Score (Goodness of fit)  
  
Results:  
| Model | RMSE | R² Score |  
|---------------------------- |----------- |--------------|  
| Linear Regression | 9409.32 | 0.758947 |

# 9. Results & Insights

- Smoking status is the strongest factor influencing charges  
- BMI also has a direct impact on higher charges

# 10. Conclusion

The project successfully developed a machine learning model to predict insurance charges. Results can help insurers price policies fairly and anticipate risk.

# 11. Future Work

- Incorporate additional health-related features (e.g., exercise, medical history)  
- Experiment with deep learning models  
- Deploy the model via a web app for real-time premium prediction

# 12. References

- Dataset source  
- Scikit-learn documentation  
- Research articles model building