# Medical Insurance Cost Prediction

A project to predict medical insurance charges using demographic and health-related data.

## Prepared By:

1. OM PRAKASH MAHATO
2. RAHUL KUMAR LAL
3. MD EBAD
4. KARAN KUMAR NONIA

## Supervisor:

Dr. Arnab Chakraborty Sir

# Table of Contents

1. Introduction  
2. Dataset Overview  
3. Data Visualization and Analysis  
4. Data Preprocessing  
5. Model Selection and Training  
6. Model Evaluation  
7. Prediction Example  
8. Conclusion  
9. References

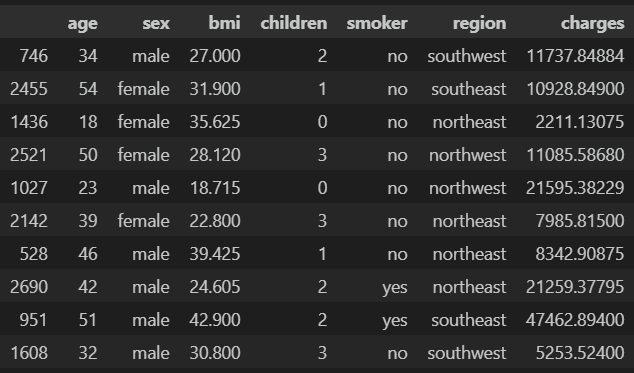
# 1. Introduction

The goal of this project is to predict medical insurance costs based on various factors such as age, BMI, smoking habits, etc.   
This predictive analysis is beneficial for insurance companies and individuals to understand factors influencing insurance costs.

# 2. Dataset Overview

The dataset used in this project contains 1338 records with 7 features:  
1. Age  
2. Sex  
3. BMI  
4. Number of Children  
5. Smoking Status  
6. Region  
7. Charges (Target variable)

A preview of the dataset is provided below:



# 3. Data Visualization and Analysis

Various visualizations were performed to understand the distribution of features and their relationships:  
1. Age Distribution (Histogram)  
2. BMI Distribution (Histogram)  
3. Correlation Heatmap

# 4. Data Preprocessing

The preprocessing steps included:  
- Encoding categorical variables (sex, smoker, region)  
- Splitting data into training (80%) and testing (20%) subsets  
- Normalizing data (if necessary)

# 5. Model Selection and Training

A Linear Regression model was used due to its simplicity and interpretability. The model was trained on the training data   
and evaluated using the R² score.

# 6. Model Evaluation

The R² scores for the model were:  
- Training data: 0.7515  
- Testing data: 0.7447  
  
This indicates a consistent and reliable performance.

# 7. Prediction Example

A sample prediction was made using the following input:  
- Age: 37  
- Sex: Female  
- BMI: 30.8  
- Children: 2  
- Smoker: No  
- Region: Southeast  
  
Predicted insurance cost: $8102.13

# 8. Conclusion

This project successfully predicted medical insurance costs with an accuracy of approximately 74%. Key insights included the   
significant impact of smoking status and BMI on charges. Future improvements could involve testing advanced models such as   
Random Forests and exploring feature engineering techniques.

# 9. References

1. Dataset: insurance.csv  
2. Libraries used: Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn