Guided by - Radha Deoghare Ma’am

Medical Cost Prediction

The aim of this analysis is to predict the medical expense based on the patients'information. The dataset used for this analysis is Insurance dataset. The dataset contains 1380 observations and 7 variables.

Project Definition

Goal

Predict individual medical expenses based on demographic and lifestyle factors using machine learning.

Problem Statement

Address the challenge of forecasting medical costs accurately to help insurers and providers set rates and manage risks

Objectives

Data Analysis:

Identify influential factors affecting medical costs

Model Building:

Develop models like Linear Regression, Decision Tree, and Random Forest.

Evaluation:

Use MAE, MSE, RMSE, and R^2 metrics to assess model accuracy.

Insights:

Analyze factors like smoking, BMI, and age to improve pricing strategies.

Data Collection and Preparation

Source:

Kaggle's "Insurance" dataset (1,380 records, 7 columns).

Attributes:

Includes age, gender, BMI, number of children, smoking status, region, and medical charges (target variable)

Data Preparation

Pre-processed data to ensure consistency and readiness for analysis

Exploratory Data Analysis (EDA)

Age, BMI, and Charges

Most patients aged 20-60, BMI range between 25-40; expenses skewed towards lower values.

Gender & Region

Balanced gender distribution; similar counts across U.S. regions.

Smoker Analysis

Higher expenses for smokers, with strong correlation to medical charges.

Correlation Insights:

- **Smoking** and **Age** have significant impacts on expenses.
- **BMI** moderately influences costs, while **children** and **region** show minimal impact.

Modeling and Machine Learning

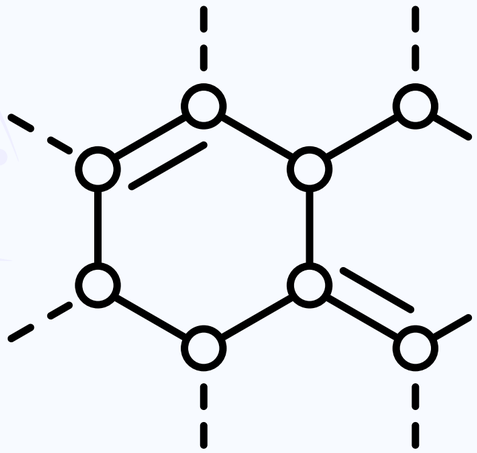
Goal:	Build predictive models for medical costs
Models Used:	Linear Regression, Logical Regression, Decision Tree, and Random Forest.
Evaluation Metrics	Assessed models based on error metrics (MAE, MSE, RMSE) and R^2 to determine accuracy.

Random Forest Regressor is best for our project

REASON

It Reduces overfitting and increases model stability by averaging across many trees.

Insights and Conclusions



- **Impact of Smoking**: Significant cost increase associated with smoking.
- **Age and BMI:** Older and higher-BMI individuals tend to have higher expenses
- **Deployment**: Model is ready for real-world application, supporting insurance and healthcare pricing.

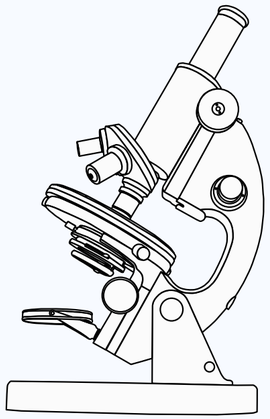
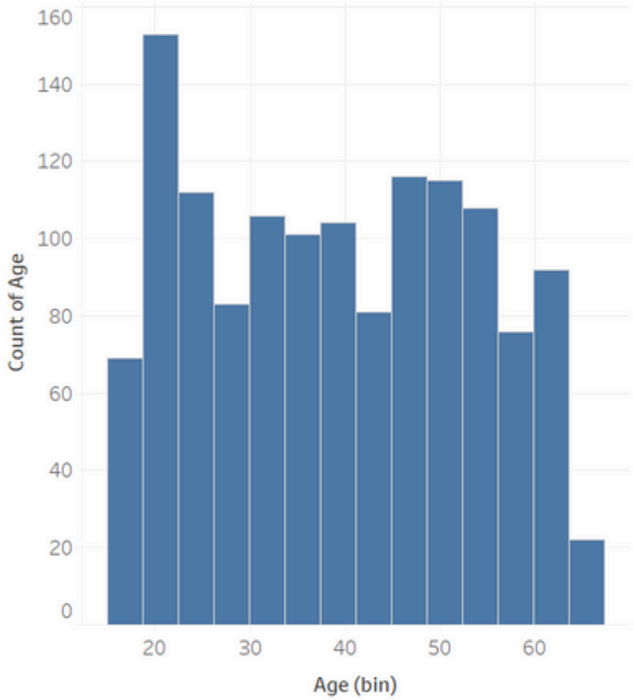


Tableau Dashboard

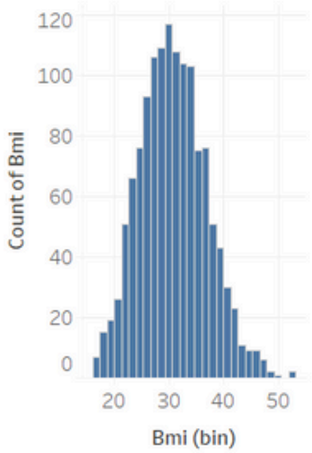
Age Distribution Analysis

This chart shows the age distribution of patients, primarily between ages 20 and 50. This information is valuable for understanding the age groups most likely to incur medical costs



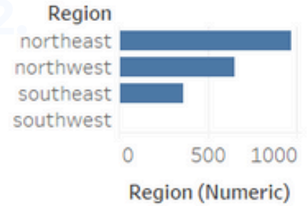
BMI Distribution Analysis

The BMI distribution is concentrated between 25 and 40, indicating a tendency toward higher BMIs. This may help assess BMI's role in medical costs.



Regional Distribution of Patients

Patients are distributed fairly evenly across all regions, with the Northwest region showing a slight concentration



THANK YOU