

Project Report
Machine Learning 1
Harini R E24AI012

Project Name: Medical Insurance cost prediction

Medical Insurance Cost Prediction – Detailed Description

Medical Insurance Cost Prediction is a **machine learning application** that estimates the **expected medical insurance charges** for an individual based on personal, lifestyle, and health-related factors. Insurance companies use such predictive models to **set fair premiums**, assess risk, and improve financial planning, while individuals can better understand how their attributes influence insurance costs.

1. Problem Overview

Health insurance charges vary widely among individuals due to differences in:

- Age
- Health condition
- Lifestyle habits
- Family size

Manually estimating insurance costs is inefficient and error-prone. Hence, **data-driven predictive models** are used to accurately forecast medical insurance expenses.

2. Objective

The primary objective is to **predict insurance charges (cost)** using historical data and machine learning techniques.

Input: insurance.csv data set

Output: Predicted medical insurance cost (in currency units)

3. Dataset Description

A commonly used dataset for this problem contains the following features:

Feature	Description
age	Age of the individual
sex	Gender (male/female)

Feature	Description
bmi	Body Mass Index
children	Number of dependents
smoker	Smoking status (yes/no)
region	Residential area
charges	Medical insurance cost (target variable)

4. Data Preprocessing

Before model training, the dataset undergoes preprocessing:

A. Handling Categorical Data

B. Convert sex, smoker, and region into numerical values using:

- a. Label Encoding
- b. One-Hot Encoding
- c. Train-Test Split

C. Dataset split into:

- a. **80% training data**
- b. **20% testing data**

5. Machine Learning Models Used

Several regression models can be applied:

a. Linear Regression

- Simple and interpretable
- Assumes linear relationship

b. Ridge Regression

- Captures non-linear patterns

6. Results and Insights

Key insights from trained models:

- **Smoking status** is the strongest factor influencing cost
- **Age and BMI** significantly impact premiums
- Individuals with more children tend to have slightly higher charges

7. Applications

- **Insurance companies:** Premium calculation, risk assessment
- **Healthcare analytics:** Cost forecasting
- **Individuals:** Understanding cost drivers
- **Policy design:** Creating fair and personalized insurance plans

8. Deployment

The trained model can be deployed using:

- **Streamlit**
- User inputs details → Model predicts insurance cost
- Real-time interactive web application

Conclusion

Medical Insurance Cost Prediction demonstrates how machine learning regression models can be effectively used to estimate healthcare expenses. It enhances transparency, reduces human bias, and enables data-driven decision-making in the insurance sector.