Medical Insurance Cost Prediction - Documentation

1. Project Overview

This project aims to predict medical insurance charges based on demographic and health-related attributes like age, BMI, smoking habits, etc.

It utilizes regression models and interactive data visualizations.

2. Dataset Description

The dataset contains the following features:

- age
- sex
- bmi
- children
- smoker
- region
- charges (target)

Data is cleaned to remove inconsistencies and ensure accurate feature types.

3. Data Preprocessing & Cleaning

Performed null checks, outlier detection, encoding of categorical features, and feature scaling.

Children column was converted to a dropdown input in the app for better UX.

4. Exploratory Data Analysis (EDA)

EDA includes:

- Univariate, Bivariate, Multivariate analysis
- Outlier detection

- Correlation analysis

Visualizations are provided through Streamlit using Seaborn and Matplotlib.

5. Model Development

Multiple regression models were tested. The best model by R^2 was a Decision Tree.

MLflow was used to track experiments and register the model.

6. Streamlit App

The app includes:

- Introduction page with context and image
- EDA section with dropdown questions and dynamic charts
- Prediction section with user input form and predicted insurance charges

7. Business Insights

Key insights:

- Smokers have significantly higher charges.
- Age and BMI positively correlate with cost.
- Obese smokers are the highest cost group.

8. Deliverables

- Python scripts for data handling and model training
- Cleaned CSV dataset
- MLflow-tracked regression model
- Streamlit app

- Project documentation (this PDF)

9. Requirements

All packages listed in requirements.txt including:

- pandas
- numpy
- matplotlib
- seaborn
- scikit-learn
- xgboost
- mlflow
- streamlit
- fpdf