# **Project Report: Medical Insurance Cost Prediction**

## **🔹 Step 1: Exploratory Data Analysis (EDA)**

**Objective:** Understand the dataset, uncover patterns, and prepare features for modeling.

* **Dataset:** Medical insurance dataset with features – age, gender, BMI, children, smoking status, region, and medical cost.
* **EDA Activities:**
  + Checked for missing values and outliers.
  + Visualized impact of age, BMI, smoking, and children on insurance costs.
  + Created an additional feature obese (BMI > 30 → True, else False).
  + Encoded categorical variables (sex, smoker, region, obese) using OneHotEncoding.
  + Standardized numerical variables where required.

📊 **Key Insights from EDA:**

* Smoking has the **highest impact** on medical costs.
* Higher BMI (obesity) increases costs significantly.
* Age shows a positive correlation with insurance cost.
* Regional differences exist, but are less significant.

## **🔹 Step 2: Model Building and Evaluation**

**Goal:** Predict insurance cost using regression models.

* **Models Tested:**
  + Linear Regression
  + Ridge Regression
  + Lasso Regression
  + Decision Tree Regressor
  + Random Forest Regressor
  + Gradient Boosting Regressor
  + XGBoost Regressor
* **Evaluation Metrics:**
  + R² Score
  + RMSE (Root Mean Squared Error)
  + MAE (Mean Absolute Error)

📈 **Result:**

* **XGBoost Regressor** performed best with the highest R² and lowest RMSE.
* Model retrained on the entire dataset and saved as: pickle file

## **🔹 Step 3: MLflow Integration**

**Objective:** Track experiments, hyperparameters, and models.

* MLflow logging included:
  + Model parameters
  + Evaluation metrics (R², RMSE, MAE)
  + Model artifacts (saved .pkl files)
* Registered the **best model (XGBoost)** in MLflow Model Registry for reproducibility and deployment.

## **🔹 Step 4: Streamlit App Development**

**Objective:** Build an interactive web app for visualization and predictions.

**Features of the Streamlit App:**

1. **Sidebar Navigation:**
   1. Project Introduction
   2. Visualizations
   3. Insurance Cost Prediction
   4. About User
2. **Visualizations Section:**
   1. Dropdown menu with EDA questions.
   2. When a user selects a question, the corresponding visualization image (from IMAGES/ folder) is displayed.
   3. Used **image paths instead of full EDA code** to improve performance.
3. **Prediction Section:**
   1. User inputs: Age, Gender, BMI, Smoking Status, Children, Region.
   2. Prediction generated using saved insurance\_cost\_model.pkl.
   3. Confidence interval of ±10% displayed along with prediction.

**Conclusion:**

* Successfully built an **end-to-end Machine Learning project**.
* The app provides **visual insights** and **real-time insurance cost predictions**.
* MLflow ensures **reproducibility, tracking, and model registry**.