

Project Title	Medical Insurance Cost Prediction
Skills take away From This Project	Python, Streamlit, Machine Learning,EDA, Data Analysis, Data Preprocessing, MLflow
Domain	Healthcare and Insurance

Problem Statement

Build an end-to-end regression project to predict individual medical insurance costs based on factors such as age, gender, BMI, smoking status, and number of dependents. Utilize the provided dataset to clean and engineer features, train multiple regression models, and deploy the best model in a Streamlit application. The app should allow users to input personal health and demographic details to estimate their medical insurance costs.

Business Use Cases

- Assisting insurance companies in determining personalized insurance premiums.
- Helping individuals plan and compare medical insurance policies based on their profile.
- Supporting healthcare consultants in estimating out-of-pocket medical costs.
- Providing cost transparency and increasing financial awareness among policyholders.

Approach

Step 1: Data Preprocessing

- Load the dataset.
- Clean the data by checking for missing, inconsistent, or duplicate records.
- Encode categorical variables (e.g., gender, smoker, region).
- Perform feature engineering such as BMI classification or interaction terms.

Step 2: Medical Insurance Cost Prediction

- Perform Exploratory Data Analysis (EDA) to identify key factors influencing insurance charges.
- Train at least 5 regression models including Linear Regression, Random Forest, and XGBoost.
- Evaluate model performance using metrics such as RMSE, MAE, and R-squared.

Integrate MLflow:

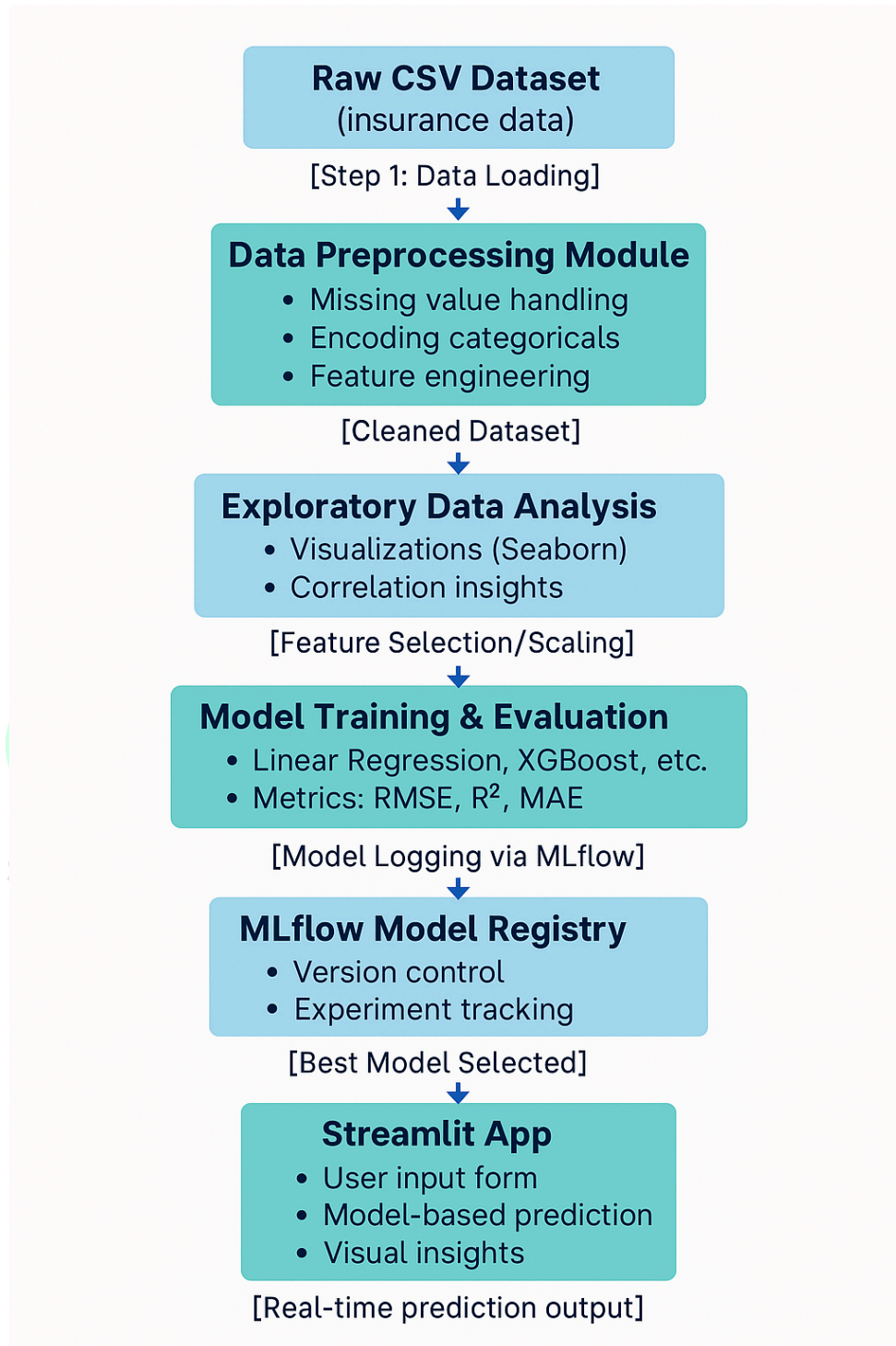
- Log experiments and metrics for each regression model using MLflow.
- Track hyperparameters, evaluation metrics, and trained model artifacts.
- Register the best performing model in MLflow's model registry.

Step 3: Streamlit App Development

Build an interactive Streamlit app that:

- Displays visual insights from EDA (e.g., impact of smoking or age on costs).
- Accepts user inputs: age, gender, BMI, smoking status, children, and region.
- Predicts the estimated medical insurance cost using the trained model.
- Optionally shows confidence intervals or error margins.

Data flow and Architecture



Dataset

Dataset Link: [medical_insurance.csv](#)

Dataset Description

Features Included:

- **age**: Age of the primary beneficiary
- **sex**: Gender (male/female)
- **bmi**: Body mass index
- **children**: Number of dependents covered by health insurance
- **smoker**: Whether the individual is a smoker
- **region**: Residential area (northeast, northwest, southeast, southwest)
- **charges**: Individual medical costs billed by health insurance (target variable)

Exploratory Data Analysis (EDA)

EDA Questions for Analysis

1. Univariate Analysis (Single Variable):

- What is the distribution of medical insurance charges?
- What is the age distribution of the individuals?
- How many people are smokers vs non-smokers?
- What is the average BMI in the dataset?
- Which regions have the most number of policyholders?

2. Bivariate Analysis (Two Variables):

- How do charges vary with age?
- Is there a difference in average charges between smokers and non-smokers?

- Does BMI impact insurance charges?
- Do men or women pay more on average?
- Is there a correlation between the number of children and the insurance charges?

3. Multivariate Analysis (More than Two Variables):

- How does smoking status combined with age affect medical charges?
- What is the impact of gender and region on charges for smokers?
- How do age, BMI, and smoking status together affect insurance cost?
- Do obese smokers (BMI > 30) pay significantly higher than non-obese non-smokers?

4. Outlier Detection:

- Are there outliers in the **charges** column? Who are the individuals paying the highest costs?
- Are there extreme BMI values that could skew predictions?

5. Correlation Analysis:

- What is the correlation between numeric features like age, BMI, number of children, and charges?
- Which features have the strongest correlation with the target variable (**charges**)?

Results

- Cleaned and processed dataset ready for analysis.
- Trained and validated regression models with optimal performance.

- A responsive Streamlit web app to display EDA and visualizations and to estimate medical insurance costs.
- Tracked model development lifecycle using MLflow.

Project Evaluation Metrics

- Completeness and accuracy of data preprocessing.
- Performance of regression models (RMSE, MAE, R-squared).
- Ease of use and reliability of the Streamlit interface.
- Clarity and value of EDA visualizations.
- Effectiveness of MLflow logging and model management.

Technical Tags:

Python, Data Cleaning, EDA, Feature Engineering, Machine Learning, Regression, Streamlit, MLflow

Deliverables:


- Python scripts for data cleaning, feature engineering, model training, and MLflow integration.
- A clean and transformed dataset in CSV format.
- Regression models for cost prediction registered and tracked via MLflow.
- A functional Streamlit web app with prediction and visualization features.
- Complete documentation covering the methodology, analysis, and business insights.

Timeline:

Check your email for submission deadlines related to this project.

References:

Project Live Evaluation Metrics	Project Live Evaluation
EDA Guide	Exploratory Data Analysis (EDA) Guide
Capstone Explanation Guideline	Capstone Explanation Guideline
GitHub Reference	How to Use GitHub.pptx
Streamlit recordings (English)	Special session for STREAMLIT(11/...
Streamlit recordings (Tamil)	
Streamlit documentation	Install Streamlit
ML FLOW Tutorial 1	ML FLOW 1
ML FLOW Tutorial 2	ML FLOW 2
MLflow DOCUMENTATION:	Getting Started with MLflow

Project Orientation (English)	
Project Orientation (Tamil)	 Project Orientation Session :Medical In...

PROJECT DOUBT CLARIFICATION SESSION (PROJECT AND CLASS DOUBTS)

About Session: The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.

Note: Book the slot at least before 12:00 Pm on the same day

Timing: Monday-Saturday (4:00PM to 5:00PM)

Booking link : <https://forms.gle/XC553oSbMJ2Gcfug9>

ss

LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)

About Session: The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.

Note: This form will Open only on Saturday (after 2 PM) and Sunday on Every Week

Timing:

For DS and AIML

Monday-Saturday (05:30PM to 07:00PM)

Booking link : <https://forms.gle/1m2Gsro41fLtZurRA>

Evaluation Metrics : [Project Live Evaluation](#)

Project Created By	Verified By	Approved By
Vinodhini	Nehlath Harmain	Nehlath Harmain

