**Predicting Medical Insurance Costs**

This project addresses the “Predicting Insurance Costs and Analysing Key Factors” challenge. We have built three model pipelines to predict medical charges and have provided a full analysis of our feature engineering and model performance.

**Project Structure**

|- ForestTrees

|-README

|-insurance\_cost\_prediction

|-data

|-insurance.csv

|-models

|-lasso\_pipeline.joblib

|-linear\_regression\_pca\_pipeline.joblib

|-xgboost\_final\_model.joblib

|-notebooks

|-ForestTrees\_EDA.ipynb

|-predict.ipynb

|-train.ipynb

|-src

|-predict.py

|-preprocessing.py

|-train.py

**Setup & Installation**

1. Download the ForestTrees Google Drive Folder provided (<https://drive.google.com/drive/folders/1EBwQOkn2vH0wYOQAL9X4N3UVsjmg9I8o?usp=sharing>)
2. Upload to your own Google Drive under My Drive
3. Make sure the ForestTrees folder that is under My Drive has the exact same folder structure listed above

**How to Run**

We have provided our pre-trained models for immediate results (under models)

1. View Our Final Predictions (Recommended)
   1. In predict.ipynb, it will load our saved models (.joblib files) and run them on the test data to reproduce our final error metrics and comparison graphs. More details are documented in the Colab notebook.
2. Retrain All Models (Optional)
   1. The script will re-train all three models (Lasso, PCA + Linear Regression, XGBoost) from scratch using the raw training data.
   2. For a better experience, it is recommended to do this in train.ipynb. More details are documented in the notebook.

**Key Analysis & Justification:**

All of our research, visualisations, and feature engineering decisions are documented in our analysis notebook (ForestTrees\_EDA.ipynb).

**Note:**

preprocessing.py under src folder contains the pipeline built for feature engineering and the function to do train-test-split. The other Python scripts under the same directory are mainly for reference and were generated from their Colab notebooks.