# Project report: Predicting Medical Insurance Charges

The goal of this project was to predict individual medical insurance charges based on demographic and lifestyle factors. The dataset used is the well-known Medical Cost Personal Dataset, which contains information on individuals insurance charges along with features such as age, sex, body mass index (BMI), number of children, smoking status, and geographic region. These features were selected because they are commonly associated with healthcare costs and are widely used in actuarial models. I focused on all the given features except the target variable charges. The target variable is the insurance cost to be predicted.

For modelling approach, the relationship between features and insurance charges, I employed linear regression as the baseline model. Given that some features were categorical (e.g., sex, smoker, region), I encoded these into numerical values. I also scaled the features using StandardScaler to normalise the data and improve model stability. Noticing that the relationship between features and charges might be nonlinear, I extended the approach to include polynomial features. This allowed the model to capture interaction and nonlinear effects among the predictors. To select the best polynomial degree, I applied GridSearchCV with cross-validation, systematically evaluating degrees from 1 to 5. This approach automated hyperparameter tuning and helped balance bias and variance.

Then, I evaluated model performance using the Mean Absolute Error (MAE), Mean Squared Error (MSE), and $R^2$ score, which are standard metrics in regression tasks. . The model achieved an MAE of approximately $2934 and an $R^2$ score of 0.83 on the training set, and an MAE of approximately $2751 and an $R^2$ score of 0.87 on the testing set. These results indicate that the model not only fits the training data well but also generalises effectively to testing data.

One challenge I faced was handling categorical variables and ensuring consistent encoding without introducing missing or unexpected values. Initially, the presence of NaNs in the sex column required careful preprocessing, including cleaning string values and confirming data types.