# **FLOW –**

# **Predicting Health Insurance Costs: A Machine Learning Approach**

◇ Health Insurance Cost Prediction Introduction-

**▹Health insurance cost prediction is a critical task in the healthcare industry, helping individuals and organizations estimate future healthcare expenses. Factors such as age, gender, smoking status, location, and number of children play a significant role in determining health insurance costs.**

**▹Age is a key factor in health insurance cost prediction, as older individuals tend to have higher healthcare needs and therefore higher insurance costs. Gender also plays a role, with insurance companies often charging different rates for males and females based on actuarial data.**

**▹Smoking status is another important factor, as smokers are at higher risk for various health conditions and may therefore face higher insurance premiums. Location can also impact insurance costs, as healthcare costs vary by region due to differences in healthcare infrastructure, provider costs, and population health.**

**▹The number of children in a family can also affect health insurance costs, as larger families may require more healthcare services and therefore face higher insurance premiums.**

**▹Predicting health insurance costs accurately is crucial for insurance companies to properly assess risk and set premiums, as well as for individuals and families to plan and budget for healthcare expenses. By leveraging machine learning algorithms and analysing historical data, we can develop models to predict health insurance costs based on these factors, ultimately improving cost estimation accuracy, and helping individuals make informed decisions about their healthcare coverage**.

**◇Importance of proposal**

**▹How predicting a health insurance cost accurately can affect the health care/insurance field?**

**Predicting health insurance costs accurately is crucial in today's world for several reasons:**

Accurate predictions help individuals and families plan and budget for their healthcare expenses, reducing financial uncertainty and stress. Insurance companies can better allocate resources and set premiums based on accurate predictions, leading to more efficient use of healthcare resources. Accurate predictions help insurance companies assess risk more effectively, leading to fairer premiums and better coverage options for individuals. By understanding future healthcare costs, individuals can make informed decisions about their health insurance coverage, ensuring they have access to the care they need. Governments and policymakers can use accurate predictions to inform healthcare policy decisions and allocate resources effectively.

**◇Significance of proposed method**

**▹How your proposed method can be helpful if required in the future for any other type of insurance?**

Machine learning models used in health insurance cost prediction can be adapted for other types of insurance, like auto or home insurance, due to similarities in data and prediction goals.Insights from health insurance cost analysis, like the impact of demographic factors, can apply to other insurance types, indicating broader trends in cost factors.Novel methodologies in insurance pricing can be applied broadly, enhancing pricing accuracy and risk assessment in different insurance contexts.These methodologies can lead to more accurate.

UNDERSTANDING THE DATASET AND FEATURES

*Most of the countries finalize health insurance costs based on many factors such as age, number of people in families, etc. What should be the actual health insurance price for an individual or a family is an issue for many companies. Hence, one insurance company hired you as a data scientist to predict the health insurance cost for possible future customers. They have already collected samples required to perform all data analysis and machine learning tasks. Your task is to perform all data analysis steps and finally create a machine learning model which can predict the health insurance cost.*

Data Description:

*The dataset contains information about individuals and families andtheir health insurance costs.It includes variables such as age, sex, BMI, number of children, smoking status, residential region, and medical costs.*Feature Analysis:1.Numerical Features:Age: *The age of the primary beneficiary.*BMI: *Body Mass Index, indicating body weight relative to height.*Children*: Number of children covered by health insurance.*Charges: *Individual medical costs billed by health insurance.*

*2.*Categorical Features:Sex: *Gender of the insurance contractor (female or male).*Smoker: *Smoking status (smoker or non-smoker).*Region: *Residential area in the US (northeast, southeast, southwest, northwest).* Data Quality: *Assess the overall quality of the dataset, including the consistency of data entry and any potential errors or anomalies.*

Target Variable: *The target variable is the "charges" column, representing individual medical costs billed by health insurance.*

*STEPS: -*

Importing Libraries*: You imported essential libraries like pandas and matplotlib.*

Data Loading and Checking*: You loaded the dataset and checked for duplicate columns.*

*.*Handling Missing Values*: Identified and extracted rows with missing values in 'health insurance price', 'age', and 'BMI'.*

Feature Engineering*: Created two additional columns 'H category' and 'BMI category' to categorize age and BMI.*

Exploratory Data Analysis (EDA):*Explored the distribution of gender, age categories, and BMI categories.Noted a high concentration of people aged 18 to 23 and a high number of obese or overweight individuals.Observed a normal distribution in the dataset.Insights:Noticed a higher number of non-smokers compared to smokers.Examined the distribution of people in different regions, which was balanced.Identified most health insurance costs to be under $10,000.Observed a weak positive correlation between health insurance prices and BMI.*

Data Preprocessing:*Encoded categorical columns (gender, location, smoking status) to numerical values.Imputed missing values in the age and BMI columns using KNN imputation.Split the dataset into training and testing sets (80/20 split).*

Model Development:*Developed six models: Linear Regression, Decision Tree Regression, Random Forest Regressor, KNN Regressor, Support Vector Regressor, and XG Boost Regressor.Selected Random Forest Regressor based on low Mean Squared Error and high R-squared score (0.85).*

Model Evaluation:*Used the selected model to predict health insurance costs for previously null values.Checked the model's performance by comparing predicted and actual values for existing entries.*

Saving and GUI of Model:

*Training the model on the entire dataset and saving it using job lib for future use.*

*Testing the saved model on specific rows of the dataset to check its performance.*

*Adding more entries to the input data dictionary to get predicted prices for multiple individuals.*

*Incorporating a graphical user interface (GUI) using (Tk inter) to interactively input data and receive insurance price predictions.*

*The GUI allows users to input age, gender, BMI, number of children, smoking status, and location, and then predicts the health insurance cost based on these inputs using the trained machine learning model. Users can click the "Predict" button to get the predicted insurance cost displayed on the GUI. This interactive interface provides a user-friendly way to use the predictive model you've developed.*