Dataset Link - *https://www.kaggle.com/datasets/mirichoi0218/insurance/data*

Group Members - Swate Patel, Aditiya Jain, Roman Marach, Adit Nuwal, Zaid Hasan

Description of Dataset

- **Source:** The dataset is uploaded on Kaggle by a user Miri Choi and is titled "Insurance". The dataset is a cleaned and formatted version, adapted from the book Machine Learning with R by Brett Lantz. The dataset is about insurance costs and can be found in the public domain. The original source of the data isn't directly linked to an organization but can be found on Kaggle .
- **Scope and Size:** The dataset contains 1,338 rows where each row represents an individual insurance policyholder. There are 7 columns in the dataset - age, sex, bmi, children, smoker, region and charges. The data is well-rounded as it includes demographic features (like age and gender), lifestyle factors (smoking status), geographic region, and the main variable of interest, charges, which indicates medical costs billed to the insurer. The dataset offers a manageable size for exploratory data analysis while providing enough features to develop meaningful predictive models.
- **Quality and Completeness:** The dataset looks well-prepared, with no missing values or data quality issues. The columns are structured properly, and categorical variables are already encoded as text which makes it easy to transform for analysis. I have checked for potential outliers in the charges column since extreme values could affect the model's performance.

Target Variable:

- Cost of Insurance, Number in Dollars, Per month or Per year. This will represent the total charge for an individual's insurance premium.

Significance:

- **Healthcare Spending Trends:** Analyzing medical costs helps identify how healthcare spending changes over time, ensuring that insurance providers understand which areas need more attention. This knowledge allows for more accurate premium setting based on actual patient needs and trends.
- **Insurance and Policy Implications:** By gaining a comprehensive understanding of medical costs, insurance companies can create tailored plans that accurately reflect the risks and needs of individuals. This leads to fairer premiums, ensuring that each individual pays what they truly need, which is the primary goal of this project.
- **Financial hardship:** Medical costs can significantly burden families, making it crucial for insurance to reflect these realities. By aligning premiums with actual medical expenses, individuals can avoid excessive financial strain, ensuring they are not overpaying for coverage.
- **Alignment with Project Goals:** The focus on medical costs directly supports our aim of ensuring individuals receive the right premium for their healthcare needs. This project's insights will guide providers in establishing fair pricing, ultimately creating a more equitable healthcare system where everyone pays appropriately based on their unique circumstances.

Features Identified:

The chosen dataset includes several key features that are expected to influence the cost of insurance premiums. These features are:

- **Age**: Represents the individual's age. Age is a critical factor in healthcare costs, as medical expenses generally increase with age.

- **BMI** (Body Mass Index): BMI is a measure of body fat based on an individual's weight and height. It is linked with various medical conditions affecting insurance costs.
- **Smoking Status**: Indicates whether an individual smokes or not. Smoking is associated with increased health risks, leading to higher medical expenses.
- **Region**: Represents the geographic location of the individual. Healthcare costs can vary by region due to differences in medical costs and regulations.
- **Gender**: Refers to the individual's gender. Gender can play a significant role in healthcare usage patterns and associated costs, as some medical conditions and preventive care measures vary by gender.

Relevance:

The rationale behind selecting these features is based on domain knowledge and statistical analysis:

- **Age**: Older individuals typically have higher healthcare utilization and costs, making age a key predictor of insurance premiums.
- **BMI**: Higher BMI values correlate with increased health risks such as cardiovascular diseases, diabetes, and hypertension, raising medical costs.
- **Smoking Status**: Smoking significantly increases the risk of chronic conditions, which drives up medical expenses and, consequently, insurance premiums.
- **Region**: Different regions have varying healthcare policies and costs, impacting insurance prices.
- **Gender**: Gender affects health conditions and risks, influencing healthcare expenses. For example, women may incur higher preventive care costs (e.g., maternity and reproductive health services), while men may be at higher risk for specific chronic diseases.

Statistical analysis, such as correlation and mutual information, shows these features to be strongly correlated with the target variable (insurance cost), confirming their significance.

Potential Impact:

These features can significantly influence the insurance cost outcomes:

- **Age** is expected to have a positive impact, as older individuals usually require more medical services.
- **BMI** may show a non-linear relationship with costs, with high BMI values linked to increased medical expenses.
- **Smoking Status** will likely have a substantial positive impact due to increased medical risks among smokers.
- **Region** can reveal variations in premiums based on regional healthcare pricing differences.
- **Gender** is likely to reveal differences in healthcare costs related to gender-specific conditions and preventive care needs.

Application Domain

Our project operates within the healthcare insurance industry. This application domain is crucial for addressing the financial and operational needs of both insurance providers and policyholders in the U.S. healthcare system. The healthcare insurance sector seeks to balance the cost of healthcare services with fair pricing strategies for policyholders. By using personal and health-related data (e.g., age, BMI, smoking status, and region), our project aims to build a model that can predict insurance prices with greater alignment to individual medical costs.

Enhancing Premium Accuracy: Healthcare insurance premiums often rely on broad risk categories, which can lead to overcharging or undercharging individuals based on general assumptions. By factoring in specific variables like age, BMI, and smoking status, this project aims to create a more nuanced approach, leading to fairer premiums that reflect true healthcare expenses. This aligns with the industry's need for data-driven methods to set premiums with precision, minimizing financial discrepancies between individual costs and coverage fees.

Reducing Financial Strain: Medical expenses are often unpredictable and can be a significant financial burden. The project aims to address this issue by ensuring that premiums are closely aligned with actual healthcare usage. By predicting premiums based on personal health factors and past costs, the model supports more tailored and fair pricing structures. This helps avoid excessive charges for individuals, thereby reducing financial hardship—a core concern in the healthcare insurance domain.

By examining how individual characteristics impact medical costs, the project aids in identifying key areas of healthcare spending, **helping insurers allocate resources more effectively**. Policy Customization: The model's data-driven approach provides insurers with insights to **design policies that reflect real medical needs.** This makes premiums fairer for individuals by accurately assessing risk, resulting in a more responsive healthcare system. Equity in Premium Setting: Accurate premiums alleviate financial strain by matching premiums to individual medical expenses, **ensuring that each policyholder's unique circumstances are considered.**