**Data Science Term Project Proposal \_Team 5**

1. **Decide on a dataset for the term project** 
   1. This dataset consists of information on 500 insurance subscribers and includes 14 features for each customer. The variables encompass a variety of categorical and numerical data related to personal information, health status, and insurance details, making the dataset well-suited for real-world insurance product design and analysis.

Total number of row : 500

Total number of feature : 14

Numerical features : Age, Annual Income, Health Score, Credit score

Categorical features : Gender, Marital Status, Occupation, Smoking Status

Target variable : Premium Amount

* 1. **The reason for selecting this dataset**
     1. **Realistic topic**: Insurance premiums are a highly important factor in real financial products, and predicting how the rates vary depending on individual characteristics can be directly applied to real-life situations.
     2. **Diverse feature composition**: The dataset includes variables that reflect a variety of factors such as personal information, health indicators, credit-related factors, and other attributes. This enables comprehensive analyses such as correlation analysis, clustering, and regression modeling.
     3. **Includes missing values**: Since the dataset contains missing data, it is considered suitable for practicing the entire data analysis process, including data preprocessing, encoding, and scaling.
     4. **Applicable to classification, regression, and clustering**: The dataset is suitable for clustering by grouping customers based on their characteristics, and insurance premiums can be predicted using regression models based on these groups.

1. **Demonstrate the description of the project**
   1. **Preprocessing**
      1. **Handling missing values**: Some variables in the dataset contain missing values (e.g., Age, Income, Credit card). For numerical variables, missing values can be replaced with the mean or median, or the corresponding rows can be removed if the missing rate is too high. For categorical variables, missing values can be filled with the mode or a separate category such as null.
      2. C**ategorical encoding**: Categorical variables represented as strings are converted using one-hot encoding, allowing the model to interpret categorical data in a numerical format.
      3. **Scaling**: Since the numerical variables have different scales, they are normalized using methods such as Standard Scaler or MinMax Scaler to ensure consistency across numerical features.
   2. **Classification by factor**

Rather than using all features indiscriminately, variables with similar characteristics are grouped to form factor-based categories as follows, and a score is calculated for each group.

* + 1. **Personal Information factors** : Age, Gender, Marital Status, Education level, Occupation
    2. **Health related factors** : Smoking Status, Exercise Frequency, Health Score
    3. **Other factors** : Insurance Duration, Credit Score
    4. **Within-group normalization** : Variables within each factor group are normalized using the same scale (e.g., placing age = 20 and education level = Master's on the same basis). Group scores are then calculated using the mean or weighted average of the normalized variables
  1. **Clustering**
     1. **Input variables** : Clustering is performed using either all features or the factor-based scores as input variables.
     2. **Clustering method** : Unsupervised learning algorithms such as K-means are used to group customers based on similar attributes.
     3. **Example** : Elderly group with health concerns – high-risk customers, leading to higher insurance premiums

Middle-aged, healthy, non-smoking, high-income group – low-risk and preferred customers, resulting in lower premiums

Low-income or newly enrolled customers – limited history and lower income imply moderate to high risk, making premium estimation possible based on their characteristics

* 1. **Clustering based regression**
     1. **After clustering customers based on similar characteristics, separate regression models will be trained for each cluster to predict the Premium Amount.**
     2. T**he data is segmented by customer group, and a separate regression model is applied to each group to predict insurance premiums.**
     3. **Regression model objective** : To design a regression model that takes customer characteristics as input and predicts the Premium Amount.
  2. **Analysis** 
     1. By comparing the regression results across clusters, we plan to analyze which groups are easier to predict and which exhibit greater variance. For groups with poor predictive performance, additional modeling strategies will be considered.
     2. We plan to derive insights that can be applied to real-world insurance company decision-making by analyzing the regression coefficients or feature importances from the regression results

1. **Demonstrate the statistical description of the dataset**

An overview of the numerical variables shows that the average customer age is approximately 42 years, and the average annual income is around $44,000. However, the income variable has a wide distribution range and a high standard deviation, indicating the need for scaling. The health score has an average of 29, with significant variation in health status among customers. The premium amount averages around $900, but the maximum exceeds $4,500, indicating a skewed distribution.

Some numerical variables contain missing values, so imputation using the mean or median will be necessary during preprocessing. Most categorical variables have 2 to 5 unique values, showing a balanced distribution. However, the *Occupation* variable has about 25% missing values, which requires careful handling. Accordingly, all categorical variables will be converted into numerical format using one-hot encoding for analysis.

Overall, this dataset includes a variety of variable types and a sufficient number of observations, making it suitable for model training after preprocessing. Additionally, the clear differences in variable distributions suggest that the dataset is highly useful for insurance premium prediction and customer segmentation analysis.







