Final Project Proposal: Insurance Data Analysis

Introduction

This proposal aims to outline a comprehensive plan for the final AI project. I aim to analyze an insurance dataset available on Kaggle (Link here -> https://www.kaggle.com/datasets/mirichoi0218/insurance/data).
As it can be seen, this dataset contains Insurance charges of various anonymized patients. Understanding insurance charges in healthcare is of paramount importance, particularly in the United States, where healthcare costs are a topic of extensive discussion. The US healthcare system involves complex interactions between insurers, healthcare providers, and individuals. Insurance charges, often driven by factors like age, health habits, and regional disparities, have a profound impact on access to medical care. Therefore, exploring the determinants of insurance charges through this dataset can provide valuable insights into a critical aspect of the US healthcare landscape.

I will utilize various machine learning techniques, including K-Means Clustering, Hierarchical Clustering (both Agglomerative and Divisive), to gain insights into the factors affecting insurance charges and customer segmentation.

Dataset Content: This comprehensive dataset encompasses various attributes, including:
Age: The age of the insured individual.
Sex: The gender of the insured (e.g., male or female).
BMI (Body Mass Index): A numerical value indicating the individual's body mass index.
Children: The number of children or dependents covered by the insurance.
Smoking Status: A categorical variable denoting whether the insured is a smoker or non-smoker.
Region: The geographical region of the insured (e.g., northeast, northwest, southeast, southwest).
Insurance Charges: The key target variable, representing the individual's insurance charges.

In this dataset, I will be using encoding for the attributes with high correlation to the Insurance charges after visualization.

This project seeks to answer a series of questions, that include, among others:
"What are the primary factors influencing insurance charges?"
"Can we accurately predict insurance charges based on individual attributes?"
"Is there a significant difference in insurance charges between smokers and non-smokers?"
"Can we identify distinct customer segments based on age, BMI, and the number of children, and what are the unique correlations of these segments?"
"How do regional variations impact insurance charges?"

Machine Learning Models that I plan to use:

The project will employ the following machine-learning techniques along with a comparison between their accuracies:

Clustering

K-Means Clustering with Elbow Method

By employing K-Means Clustering with the Elbow Method, I aim to identify the ideal number of customer segments within my dataset. This will allow me to gain valuable insights into how age, BMI, and the number of children contribute to the formation of these segments. Moreover, this technique provides a structured approach to segmenting the customer base effectively.

Hierarchical Clustering

a) **Agglomerative Hierarchical Clustering**
b) **Divisive Hierarchical Clustering**

Both Agglomerative and Divisive Clustering methods provide valuable insights into the hierarchical relationships and structures present in the data. By visualizing the dendrogram, I can identify meaningful customer segments and understand how different attributes contribute to their formation. This hierarchical perspective allows for a nuanced understanding of the dataset, which complements the findings from K-Means Clustering.

Regression

In this project, regression models will be applied to predict insurance charges based on various individual attributes. Two types of regression will be used:

Linear Regression

Linear Regression: Linear regression models will be employed to establish a baseline for predicting insurance charges. This approach assumes a linear relationship between the input attributes (e.g., age, BMI, smoker) and the target variable (insurance charges). The coefficients of the linear equation will provide insights into the strength and direction of these relationships.

Decision Tree Regression

Decision tree regression will be utilized to build predictive models for insurance charges. These models segment the data into subsets based on attribute values and make predictions based on the average of target values in each subset.

The regression analysis will enable a detailed understanding of how each attribute contributes to insurance charges and provide valuable insights into the nature of these relationships.

K-Fold Cross-Validation

To ensure the robustness and reliability of the regression models employed, I will utilize **K-Fold Cross-Validation** as a key step in model evaluation. K-fold cross-validation is a widely adopted technique for assessing a model's performance by partitioning the dataset into K subsets or folds.

Model Comparison

I will also perform a comparative analysis of different machine learning models to determine their effectiveness in addressing the research questions to evaluate the predictive performance and interpretability of each technique. By systematically assessing the strengths and limitations of different clustering and regression, I aim to identify which models are best suited for specific aspects of the analysis.