Insurance Premium Prediction Project

Dataset Link:

The dataset used in the project can be found on Kaggle:

https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

Description of Project:

The project aims to build a model that predicts the insurance premium for customers based on their demographics and medical history. The dataset used for this project is the Medical Cost Personal Datasets available on Kaggle. The project includes data preprocessing, exploratory data analysis, feature selection, and model building.

Problem Statement

Dataset Information: The dataset used for this project contains information about individuals and their insurance premiums. It includes the following features:

1. Age: Age of the individual (numeric)

2. Sex: Gender of the individual (categorical: male or female)

3. BMI: Body Mass Index of the individual (numeric)

4. Children: Number of children the individual has (numeric)

5. Smoker: Whether the individual is a smoker or not (categorical: yes or no)

6. Charges: Insurance premium charged to the individual (numeric)

Background Information: Insurance companies use various factors to determine the premiums they charge their customers. These factors can include age, gender, BMI, number of children, and smoking status. Accurately predicting insurance premiums can help individuals understand the factors that influence their premiums and enable insurance companies to make more accurate pricing decisions.

The goal of this project is to develop a machine learning model that can predict insurance premiums based on the given features. By analyzing the relationships between the input features and the insurance premiums, the model will learn to make predictions for new individuals.

Insurance Premium Prediction Project

The provided code performs various data analysis and visualization tasks on the insurance dataset. It includes steps such as loading the dataset, exploring the data, visualizing feature correlations, handling missing values, transforming categorical variables, splitting the data into training and testing sets, training and evaluating different regression models, and comparing their performance.

The dataset contains information from different individuals, allowing us to analyze how various factors, such as age, gender, BMI, number of children, and smoking status, contribute to the insurance premiums. By developing a predictive model, we can estimate insurance premiums for new individuals based on their personal information.

Overall, this project aims to provide insights into the factors influencing insurance premiums and develop a model that can accurately predict premiums based on individual characteristics.

Train and evaluate the following models:

⏱ Multiple Linear Regression

⏱ Support Vector Regression

⏱ Polynomial Regression

⏱ Decision Tree Regression

⏱ Random Forest Regression

⏱ Create a function, model_summary(), to calculate and display metrics such as training and testing accuracy

⏱ Apply the model_summary() function to each trained model to obtain their performance metrics.

Model Comparison and Selection:

⏱ Compare the performance of different models based on their accuracy

⏱ Identify the model with the highest accuracy as the selected model for insurance premium prediction.

Insurance Premium Prediction Project

Conclusion:

🕐 Summarize the results and findings from the model evaluation.

🕐 Highlight the model with the highest accuracy as the recommended model for predicting insurance premiums.

🕐 Discuss the potential impact and benefits of accurate insurance premium prediction.

🕐 By following this outline, you can write the code for the Insurance Premium Prediction project step by step. Remember to comment your code for better understanding and maintainability.