Insurance Price Prediction Using Machine Learning

Course: Machine Learning
Project Type: Semester Project
Student Name: Sumair Mazhar

1. Introduction

Machine Learning (ML) plays a vital role in modern decision-making systems. In the insurance industry, predicting insurance prices accurately is essential for both companies and customers. This project focuses on building a machine learning model that predicts insurance charges based on customer-related attributes such as age, BMI, smoking status, number of children, gender, and region.

2. Problem Statement

Insurance companies often face challenges in estimating fair insurance prices due to the complexity of multiple influencing factors. Manual estimation may lead to biased or inaccurate pricing. This project aims to solve this problem by using machine learning to predict insurance charges based on historical data.

3. Objectives of the Project

The main objectives of this project are:
• To understand and analyze the insurance dataset
• To preprocess and clean the data for machine learning
• To build a regression-based machine learning model
• To evaluate the performance of the trained model

4. Dataset Description

The dataset used in this project contains information about insurance customers. Each record represents a single customer with the following attributes:

• Age – Age of the customer
• Sex – Gender of the customer
• BMI – Body Mass Index
• Children – Number of dependents
• Smoker – Smoking status (Yes/No)
• Region – Residential region
• Charges – Insurance cost (target variable)

5. Tools and Technologies Used

The following tools and technologies were used in this project:

- Python Programming Language
- Jupyter Notebook
- Pandas for data manipulation
- NumPy for numerical computations
- Matplotlib for data visualization
- Scikit-learn for machine learning modeling

6. Methodology

The project followed a systematic machine learning workflow:

1. Data Loading: The dataset was loaded using Pandas.
2. Exploratory Data Analysis (EDA): Basic statistics and visualizations were performed.
3. Data Preprocessing: Categorical variables were encoded and missing values handled.
4. Train-Test Split: The dataset was divided into training and testing sets.
5. Model Training: A regression model was trained on the training data.
6. Model Evaluation: Model performance was evaluated using appropriate metrics.

7. Model Implementation

A regression-based machine learning model was implemented to predict insurance charges. The model learns the relationship between input features and the target variable. Scikit-learn was used to train and test the model efficiently.

8. Results and Evaluation

The trained model produced reasonable predictions for insurance charges. Performance was evaluated using metrics such as mean absolute error and accuracy trends. The results indicate that factors such as smoking status and age have a strong influence on insurance prices.

9. Conclusion

This project successfully demonstrates the application of machine learning in predicting insurance prices. The model provides useful insights into the factors affecting insurance costs. Future improvements may include using advanced algorithms and larger datasets to enhance accuracy.

10. Future Scope

Future enhancements of this project may include:
- Using advanced machine learning models
- Hyperparameter tuning

• Deploying the model as a web application

11. References
• Scikit-learn Documentation
• Kaggle Insurance Dataset
• Python Official Documentation

12. Author Details
Name: Sumair Mazhar
Course: Machine Learning