Medical Insurance Cost Prediction Project Documentation

Project Overview

The Medical Insurance Cost Prediction project is a machine learning-based application designed to predict medical insurance costs based on user inputs such as age, BMI, smoking status, number of children, and region. The project leverages a Linear Regression model trained on synthetic data, and the user interface is developed using Streamlit for interactive prediction and visualization.

Key Features

1. **User-friendly Interface**:
    - Interactive sliders, dropdowns, and buttons for entering input features.
    - Real-time prediction updates with a single click.
    - Visualizations for better understanding of predictions and feature impacts.
2. **Machine Learning Model**:
    - **Linear Regression** model trained using Scikit-learn.
    - Encoded categorical variables (e.g., region and smoker status) to enable the model to handle non-numerical data.
3. **Feature Contributions**:
    - Breakdown of user inputs and their contribution to the prediction.
    - Visual analysis of the relationships between key features and insurance costs.
4. **Deployment**:
    - Deployed using **Streamlit** for a responsive, browser-based interface.

Technical Components

1. Dataset

The dataset medical_insurance.csv contains the following features:

- **age**: Age of the individual.
- **bmi**: Body Mass Index, a measure of body fat based on height and weight.
- **children**: Number of dependents.
- **smoker**: Smoking status (Yes/No).
- **region**: Geographical region (northeast, northwest, southeast, southwest).
- **charges**: Medical insurance cost (target variable).

---

2. Preprocessing Steps

- **One-Hot Encoding**: Applied to the categorical feature region to convert it into binary columns for each region.
- **Label Encoding**: Used for the smoker column, where "Yes" is encoded as 1 and "No" as 0.
- **Feature Engineering**: Combined all encoded features with numerical features into a single dataset for training.

---

3. Model Training

- **Algorithm**: Linear Regression from Scikit-learn.
- **Training Process**:
    - Features: ['age', 'bmi', 'children', 'smoker_encoded', 'region_northeast', 'region_northwest', 'region_southeast', 'region_southwest'].
    - Target Variable: charges.
    - Dataset split: 80% for training and 20% for testing.
- **Evaluation Metrics**:
    - Mean Squared Error (MSE).
    - R-squared Score (R^2).

---

4. Web Application

Streamlit is used to create an interactive web-based application. Key features of the app include:

- **Sidebar Inputs**: Allows users to input age, BMI, number of children, smoking status, and region.
- **Prediction Results**: Displays the predicted insurance cost and a breakdown of input features.
- **Visualizations**:
    - Feature importance bar chart.
    - Predicted insurance cost trends with age.

---

5. Required Libraries

To run the project, the following Python libraries are required:

- **Streamlit**: For building the interactive user interface.
- **Pandas**: For data manipulation and preprocessing.

- **NumPy**: For numerical operations.
- **Scikit-learn**: For machine learning model training and evaluation.
- **Matplotlib & Seaborn**: For creating visualizations.
- **Joblib**: For saving and loading the trained model.

Installation and Setup

1. Clone the repository or download the project files.
2. Install the required libraries:
3. pip install -r requirements.txt
4. Place the dataset medical_insurance.csv in the project directory.
5. Train the model and save it:
6. python train_model.py
7. Run the Streamlit app:
8. streamlit run app.py
9. Open the app in your browser at http://localhost:8501.

Project Workflow

1. **Data Loading**:
    - Load the dataset medical_insurance.csv using Pandas.
2. **Preprocessing**:
    - Encode categorical variables (region and smoker).
    - Combine processed features into a single dataset.
3. **Model Training**:
    - Split data into training and testing sets.
    - Train a Linear Regression model using Scikit-learn.
4. **Prediction and Evaluation**:
    - Evaluate model performance on test data using MSE and R^2.
    - Save the trained model using Joblib.
5. **Web App Interface**:
    - Accept user inputs for prediction.
    - Display prediction results and feature impact visualizations.

Usage

1. Launch the Streamlit app and provide the following inputs:
    - **Age**: Slider input for age (18–100 years).
    - **BMI**: Numeric input for BMI (10–50).
    - **Children**: Slider input for the number of children (0–5).
    - **Smoker**: Radio button for smoking status (Yes/No).
    - **Region**: Dropdown menu for region (northeast, northwest, southeast, southwest).
2. Click the **Predict** button to calculate the estimated insurance cost.
3. Explore the visualizations to analyze the effect of individual features on the prediction.

Results

The app provides:

- Predicted medical insurance costs based on user inputs.
- Insights into feature contributions to the prediction.
- Interactive visualizations for better understanding.

Conclusion

This project demonstrates how machine learning models can be effectively combined with interactive tools like Streamlit to create user-friendly applications for solving real-world problems. The Medical Insurance Cost Prediction app enables users to estimate their insurance costs and understand the factors influencing them.