Major Project

Medical Insurance Cost Prediction

Presented By - Pranjay kumar

Roll Number - Btech/15018/21

Faculty Mentor - Dr. Hazique Aetesam

Outline

- Introduction
- Problem Statement
- Motivation
- Flow chart
- Dataset Description
- Proposed Methodology
- Results and Discussions

Introduction

- The project aims to develop a **predictive model for estimating medical insurance costs** using individual data (age, sex, BMI, number of children, smoking status, and region), helping insurance companies assess risk and set premiums[1].
- The dataset consists of both categorical (sex, smoker, region) and numerical (age, BMI, children) features, with no missing values, allowing for smooth preprocessing and analysis.
- Data preprocessing includes encoding categorical variables into numerical values, splitting the data into training and testing sets (80/20), and preparing it for machine learning models such as **Linear Regression, Decision Tree Regression, and Random Forest Regression** .
- Model performance is evaluated using the **R-squared (R^2)** metric on both training and testing data to ensure the model explains the variability in medical charges and generalizes well to unseen data.

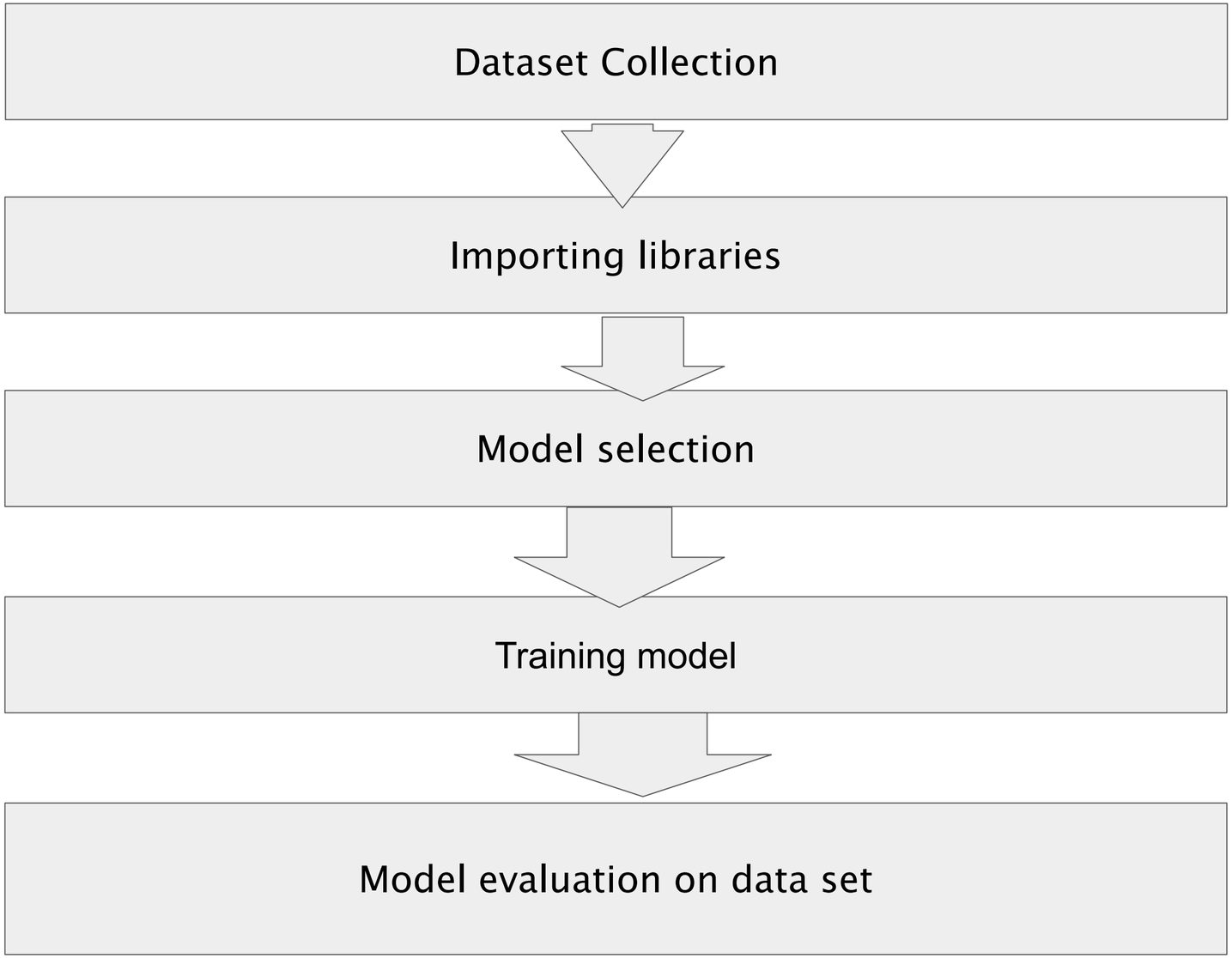
Problem Statement

- Develop a predictive model to estimate medical insurance costs based on individual features.
- Provide valuable insights for insurance companies to assess potential risks and set appropriate premiums.
- Use a dataset containing demographic and behavioral features: age, sex, BMI, number of children, smoking status, and region.
- Predict the target variable: medical charges incurred by individuals.

MOTIVATION

- The motivation for this project is to provide a **reliable system for predicting medical insurance costs**, which can help individuals estimate their **potential medical expenses and assist insurance companies in setting fair and accurate premiums .**
- Accurately forecasting medical costs is challenging due to the many factors involved, such as age, smoking status, and BMI, all of which significantly impact healthcare expenses; this project aims to leverage data-driven models to address this complexity .
- By enabling better estimation of future medical expenses, the project empowers patients to make **informed decisions about insurance claims and financial planning** and helps insurers identify high-risk customers and allocate resources more efficiently .
- The use of machine learning in this context not only improves prediction accuracy but also supports broader goals such as healthcare resource optimization, policy development and promoting transparency in insurance pricing .

FLOW CHART


Dataset Collection
Importing libraries
Model selection
Training model
Model evaluation on data set

Dataset Description

- The dataset contains **1,338 rows and 7 columns**, with each row representing an individual and each column representing a feature relevant to medical insurance costs.
- Data Types:
  - Categorical: **sex, smoker, region**
  - Numerical: **age, bmi, children, charges**
- Encoding for Modeling:
  - Sex: **male = 0, female = 1**
  - Smoker: **yes = 0, no = 1**
  - Region: **southeast = 0, southwest = 1, northeast = 2, northwest = 3**

Dataset Description(Continued)...

- Sample Data (First 5 Rows):

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.55
28	male	33.0	3	no	southeast	4449.46
33	male	22.71	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.86

Proposed Methodology

- Data preprocessing: Load the dataset, check for missing values, and encode categorical features (sex, smoker, region) into numerical values for analysis.
- Feature selection: Use age, sex, BMI, number of children, smoker status, and region as predictors; charges as the target variable.
- Data splitting: Divide the data into **training (80%) and testing (20%)** sets to evaluate model performance.
- Model training and evaluation: Train **Linear Regression, Decision Tree, and Random Forest** models; assess accuracy using **R-squared (R^2)** on both training and test sets.

Results and Discussions

On Training Data:

Model	R^2 Score (Training Data)
Linear Regression	0.7413
Decision Tree	0.9983
Random Forest	0.9758

Results and Discussions(Continued)...

- **On Testing Data:**

Model	R^2 Score (Test Data)
Linear Regression	0.7830
Decision Tree	0.7080
Random Forest	0.8629

Results and Discussions(Continued)...

- **Prediction on input values:**
- **Age,Sex,BMI,Children,Smoker,Region(31,1,25.74,0,1,0)**

Model	Predicted Insurance Cost (USD)
Linear Regression	4016.99
Decision Tree	3756.62
Random Forest	3737.32

Actual cost is **3756.6216(USD)**

Thank You

Appendix

- R-squared (R^2), also known as the **coefficient of determination**, is a statistical measure used to evaluate how well a regression model explains the variance in the dependent variable based on the independent variables.
- Interpretation:
  - R^2 values range from 0 to 1.
  - An R^2 of 1 means the model explains all the variability of the target variable.
  - An R^2 of 0 means the model explains none of the variability.