EAST WEST UNIVERSITY

Department of Computer Science and Engineering

Course Title :Statistics for Data Science.
Course Code : CSE 303.
Section : 03.
Semester : Spring25.

Submitted to
Dr. Mohammad Manzurul Islam
Assistant Professor
Department of Computer Sciences and Engineering
East West University

Submitted by

Name	Student ID
Farhatun Nahar Priya	2023-1-60-202
Afsana Akter Mim	2023-1-60-073
Nuran Farhana Prova	2023-1-60-075

Date of submission: 20 May, 2025.

Project Report

Introduction:

Medical costs are rising every year and everyone should be ready to adjust with these expenses. Calculating how much medical insurance will cost may help people in managing their finances and medical related issues more wisely.

The main goal of this project is to run a machine learning model that can predict medical insurance costs based on a person's many details. We are using a dataset that includes information like age, gender, body mass index (BMI), number of children, smoking status, and the region. These features can help to predict how much a person pays for health insurance.

When analyzing the data, we seek to understand what correlates between the information we collect and the insurance fees. After training the model can be used to predict medical costs for new people using their information. Using this system insurance companies can charge acceptable prices and people can understand what things they should do make their insurance cost more or less

For this project, first we look at the data then prepare it, train the model and check how well it works. To finish the task we count on Python and various machine learning libraries. The last model is meant to produce predictions of medical costs that are accurate and reliable.

Dataset Description:

In this project, the project's dataset includes data on individual medical insurance premiums. There are seven columns and 1,338 rows. One person's data is represented by each row, and a distinct feature about the person is provided by each column.

Both categorical and numerical data are included in the dataset:

Numbers such as age, bmi, and number of children are examples of numerical data.

Text-based values like sex, smoking status, and region are examples of categorical data.

Features List:

age (numerical): The person's age.

sex (Categorical): A person's gender, either male or female. As in the given data set the sex is given categorical so we have to mapping by putting male =1 and female =2.

bmi (numerical): is a measurement of body fat based on weight and height.

children (numerical): The total number of dependents or children that the insurance covers.

smoker (categorical): Whether or not the individual smokes. In this case we are mapping this column 1 and 0. If the person is a smoker I mean the ans is yes then it'll be replaced by 1 and if the person is a non-smoker then it'll be replaced by 0.

region (categorical): The person's home region in the United States (northeast, northwest, southeast, southwest). For this one, we replaced the region by 1 to 4. For southwest =1, southeast=2 , northwest= 3 and northeast=4.

charges (numerical): The amount we wish to forecast is the individual's medical insurance bill.

A machine learning model that predicts the charges based on the other features can be trained with the help of this dataset.

Data Preprocessing:

Before training the model we need to prepare the dataset so that it can be used properly by machine learning algorithms. This process is called data preprocessing. Below are the steps we followed:

1. Checking for Missing Values

We first checked if the dataset had any missing or null values. Fortunately, this dataset is clean and does not contain any missing values and so no rows or columns need to be removed.

2. Converting Categorical Values

Machine learning models work best with numerical values so we had to convert the categorical columns into numerical values.

- **sex**: We converted 'male' and 'female' to numbers using label encoding. For example,we use male = 1, female = 0.

- **smoker**: We converted 'yes' and 'no' into numbers. For example, we use smoker (yes) = 1 and non-smoker (no) = 0.

- **region**: Since the region column contains more than two categories (northeast, northwest, southeast, southwest) and we used label encoding to assign a unique numerical value to each region. Where southwest = 1, southeast = 2, northwest = 3, northeast = 4

3. Feature Selection

We use all the features (age, sex, bmi, children, smoker, region) because they all help in predicting the charges.

4. Splitting the Dataset

We split the data into training and testing sets. Normally we use 70–80% of the data for training and the rest of the data for testing. This helps us check how well our model performs on new and unseen data.

After these steps the dataset was ready for building and training the machine learning model.

Exploratory Data Analysis (EDA):

We have visualized the to analyze the medical cost data and notice any relationships between different features. This information covers age, sex, BMI, the number of children, if they smoke and what area they live in and each factor may determine costs for medical insurance.

Correlation Matrix:
Correlation Matrix shows the Heatmap was developed to represent how various numbers involved in the data set relate to each other. This allows us to identify features that have a close relationship. We observed that people who smoke often have to pay higher medical costs. This also implies that people who have a higher BMI are likely to spend more on their healthcare. An increase in age is linked to a slight increase in health costs.

Bar Plot:
We compared the average medical charges in different regions by using a bar plot. The plot reveals that the southeast region typically has the highest average charges. Some reasons could include the way people live, the healthcare they receive or the usual health-related issues found in that location.

Line Plot:
We created a line plot to study the relationship between individuals' age and their number of children, separating smokers and non-smokers. The data reveal that the average number of children among smokers and non-smokers is the same for each age group. Even though there are distinct differences among smokers, the general pattern for both groups does not vary significantly. This suggests that one's smoking status is not strongly related to their number of children.

Count plot:
We used a count plot to visualize the distribution of smokers and non-smokers in the dataset.
Count plots are effective for displaying the frequency of categorical variables. In this case, the plot shows that non-smokers are significantly more prevalent than smokers. This helps us understand the overall smoking behavior in the dataset and may indicate general lifestyle or health awareness trends within the population.

Histograms:
Histograms are used to visualize the distribution of a single numerical variable. In this case, we used a histogram to show the distribution of BMI.we can see that the BMI values in the data show a normal curve but are skewed to the right side. There is a spread from a BMI of 25 to 35, with the most

common BMI (mode) being 30. Using this figure, we can identify the average and wider ranges of BMI which may support recognizing patterns or warning signs in health risks.

Pie Chart:

Pie charts are used to show the proportion of different categories within a categorical variable. They are particularly useful when we want to visualize how a whole is divided into parts. In this case, we used a pie chart to illustrate the proportion of people who smoke and those who do not among individuals in the dataset.The chart clearly shows that a majority of individuals (79.5%) are non-smokers, while a smaller proportion (20.5%) are smokers. This visual representation helps in quickly understanding the distribution of smoking status and can be useful in identifying potential lifestyle-related factors when analyzing health outcomes such as stroke risk.

Box Plot:

Box plots are used to show the distribution of a numerical variable for different categories of a categorical variable. They provide information about the median, quartiles, and potential outliers within each category. The plot shows how many children one has can change their charges. There are not big differences in median costs, even so, when there are 0 to 3 children in a family, the range of expenses becomes very wide. Those with 4 or 5 children generally pay less per child, but there could be fewer families in those households.

Scatter Plot:

A scatter plot shows how two numerical variables are connected to each other. In a scatter plot, every dot represents a set of data and its location shows the values of the two factors. We used a scatter plot that included the status of smoking as an added color to analyze the connection between BMI and insurance expenses.By showing the data visually, patterns between BMI and charges are easier to detect. Higher BMI levels among smokers result in even higher healthcare charges than non-smokers.

Machine Learning Models:

In this project, we used machine learning models to see medical expenses depending on different variables including age, sex, bmi, number of children, smoker, region and charges. Linear regression was the main model applied cause it is appropriate for continuous variable prediction such as medical costs.

The reason of using Linear Regression:

We chose linear regression because since the target variable charges-is a continuous numerical value and Linear Regression is a simple model that creates a linear relationship between the input features and the target it was decided upon. It offers understanding of how every characteristic affects the expected result.

It is a popular and efficient statistical method for forecasting continuous outcomes. Here is a more detailed explanation of why it is effective for this project:

The Suitability of Linear Regression for This Project:

In this section, we'll explain why we chose Linear Regression and how it fits well with the goals of this project:

Forecasting Continuous Values:
It is a regression problem rather than a classification problem because the objective is to estimate a numerical outcome, like insurance charges. For predicting continuous variables, linear regression is ideal.

Relationship Quantification:
The relationship between independent variables (like age, bmi, and smoking status) and the dependent variable (charges) can be analyzed and quantified with the use of linear regression. This makes it possible to comprehend exactly how each factor affects the desired variable.

Interpretation and Significance of Features:

Each feature's significance is indicated by the model's coefficients. A high bmi coefficient indicates that bmi significantly affects charges.

Python Implementation Simplicity:

A straightforward interface for implementing linear regression with only a few lines of code is provided by libraries such as (LinearRegression()). It works well with workflows for machine learning.

Efficiency & Scalability:

Linear regression is scalable for handling real-world data because it works well with large datasets that contain numerical predictors.

Analysis of Trends:

The model is able to identify patterns in various populations. For instance, it might show that growing older is associated with greater health care costs.

Model Baseline for Additional Methods:

Before examining more models like polynomial regression or tree-based techniques, linear regression is frequently used as a starting point.

Linear Regression Evaluation Measures:

Measuring the average size of the mistakes in forecasts, mean absolute error (MAE) It computation is as follows: 4285.22

Measuring the average squared difference between expected and actual values, mean squared error (MSE) It comes computed as: 38364832.19

Root Mean Squared Error (RMSE): 6193.94

R^2 Score (Insert value here): 0.74

Conclusion:

In this project, we run a machine learning model using Linear Regression to predict medical insurance costs based on various features such as age, gender, BMI, number of children, smoking status and region.

After processing and training the model we evaluated its performance using common regression metrics. The results are as follows:

Mean Absolute Error (MAE): 4,285.22
This means on average the model's predictions are about 4,285 off from the actual medical costs.

Mean Squared Error (MSE): 38,364,832.19

Root Mean Squared Error (RMSE): 6,193.94
These values indicate the presence of some larger individual prediction errors which is normal in real-world cost prediction problems.

R-squared (R^2): 0.74
This score shows that the model explains 74% of the variation in medical costs based on the input features. It means the model captures most of the important patterns in the data.

From our analysis we found that smoking status, age and BMI are key factors influencing medical costs. Smokers, older people, and those with higher BMI tend to have higher insurance costs.

Overall the Linear Regression model shows reasonably good performance and provides useful insights into how different information affects medical insurance costs and expenses.