

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering



Documentation Report of the Dataset

Course No : CSE4108
Course Name : Artificial Intelligence Lab
Section : B
Lab Group : B1

Submitted To:

Mr. Md Siam Ansary
Lecturer, Dept. of CSE, AUST

Mr. Ashek Seum
Lecturer, Dept. of CSE, AUST

Submitted By

Name : Md. Sakib Irtiza
ID : 17.02.04.081

Date of Submission: September 9, 2021.

A Description of the Dataset

Name of the Dataset	Medical Insurance Dataset
File Format of the Dataset	.csv
Dimension of the Dataset	185 * 7
Number of Total Columns	7
Number of Total Rows	185
Number of Feature Columns	6
Name of Feature Columns	Age,gender,bmi,no. Of children,smoker,region
Number of Target Column(s)	1
Name of Target Columns	charges

Description

The dataset has 7 columns and 185 rows. Of 7 columns, 6 columns are feature columns and they are : age,gender,bmi,children,smoker and region. The last column is the target column which we are going to predict the value of and the name of that column is charges. Which means we are going to predict the medical insurance costs.

A short description of each columns is given below:

Name of the feature: age

Unit: Integer

Description: age of primary beneficiary. As the value of age is always integer and cannot be fraction, that is why the unit of this feature is integer.

Name of the feature: gender

Unit: Integer

Description: Gender of the insurance contractor, and they are: male and female. In the project, the categorical features of gender are converted to numerical features. Where, male is denoted as 0 and female is denoted by 1. That is why, the unit of this feature is integer.

Name of the feature: bmi

Unit : Float

Description: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9, which means this value can be a float number. That is why, the unit of this feature is float.

Name of the feature: children

Unit : Integer

Description: Number of children covered by Number of dependents and recorded by the health insurances. This feature column shows the total number of children of the insurer.

Name of the feature: smoker

Unit: Integer

Description: This feature column indicates the people if they are involved with smoking or not. In the project, the categorical features of gender are converted to numerical features. Where, the people who are not involved with smoking are denoted as 0 and the people who are involved with smoking are denoted by 1.

Name of the feature: region

Unit: Integer

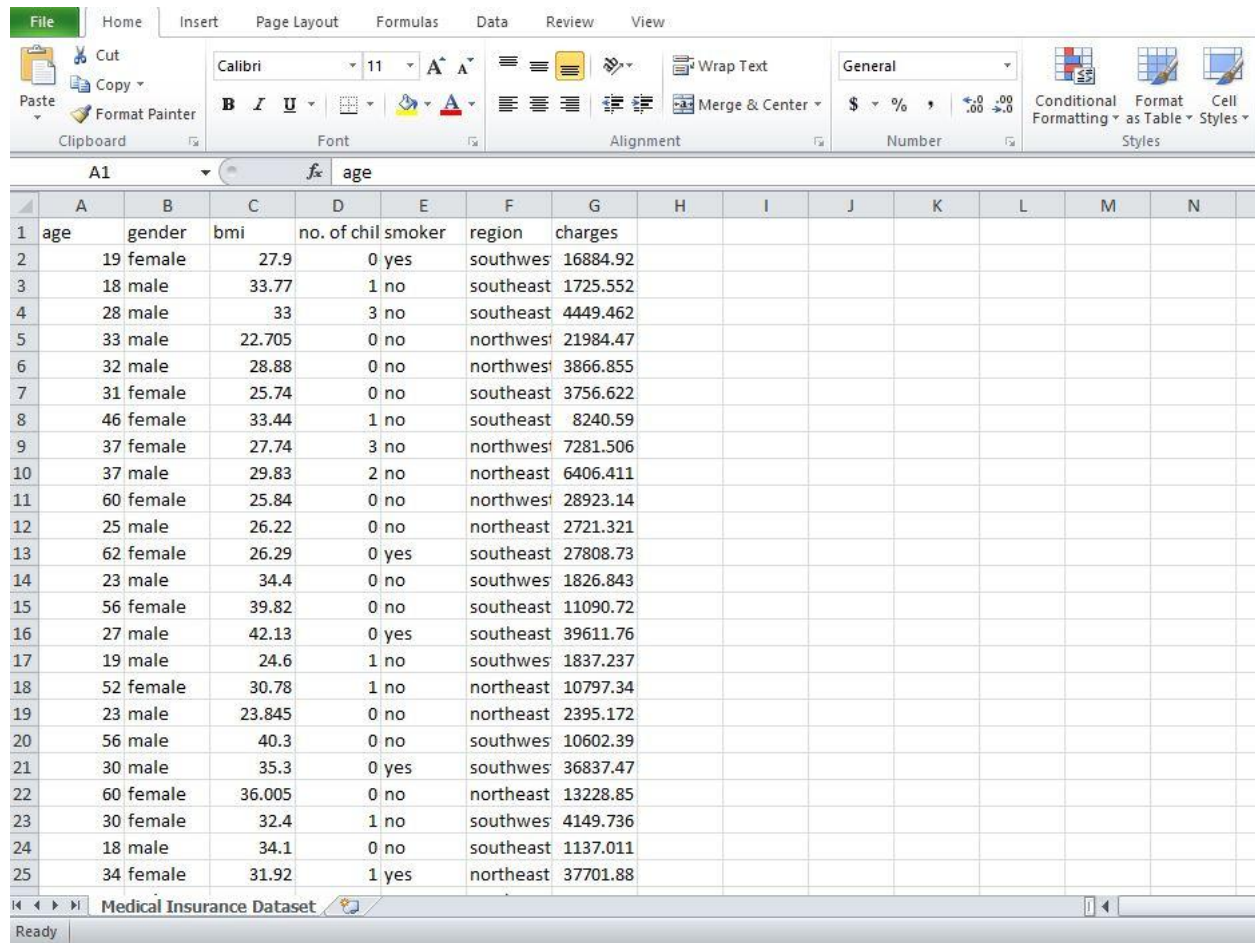
Description: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest. In the project, the categorical features of the regions of US are converted to numerical features. Where, the region "Northwest" is denoted by 0. The region "Northeast" is denoted by 1, the region "Southeast" is denoted by 2 and the region "Southwest" is denoted by 3.

Source of The Dataset :

Data are collected from the below resource :

<https://gist.github.com/meperezcuello/82a9f1c1c473d6585e750ad2e3c05a41>

A sample screenshot of the dataset used in this project is shown below:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	gender	bmi	no. of chil	smoker	region	charges							
2		19 female	27.9	0	yes	southwes	16884.92							
3		18 male	33.77	1	no	southeast	1725.552							
4		28 male	33	3	no	southeast	4449.462							
5		33 male	22.705	0	no	northwest	21984.47							
6		32 male	28.88	0	no	northwest	3866.855							
7		31 female	25.74	0	no	southeast	3756.622							
8		46 female	33.44	1	no	southeast	8240.59							
9		37 female	27.74	3	no	northwest	7281.506							
10		37 male	29.83	2	no	northeast	6406.411							
11		60 female	25.84	0	no	northwest	28923.14							
12		25 male	26.22	0	no	northeast	2721.321							
13		62 female	26.29	0	yes	southeast	27808.73							
14		23 male	34.4	0	no	southwes	1826.843							
15		56 female	39.82	0	no	southeast	11090.72							
16		27 male	42.13	0	yes	southeast	39611.76							
17		19 male	24.6	1	no	southwes	1837.237							
18		52 female	30.78	1	no	northeast	10797.34							
19		23 male	23.845	0	no	northeast	2395.172							
20		56 male	40.3	0	no	southwes	10602.39							
21		30 male	35.3	0	yes	southwes	36837.47							
22		60 female	36.005	0	no	northeast	13228.85							
23		30 female	32.4	1	no	southwes	4149.736							
24		18 male	34.1	0	no	southeast	1137.011							
25		34 female	31.92	1	yes	northeast	37701.88							

Figure: A Sample Screenshot of the Dataset