Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Project Report

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.

1 Introduction

Health insurance is a type of insurance coverage that typically pays for medical, surgical, prescription drug and sometimes dental expenses incurred by the insured.Health insurance can reimburse the insured for expenses incurred from illness or injury, or pay the care provider directly. It is often included in employer benefit packages as a means of enticing quality employees, with premiums partially covered by the employer but often also deducted from employee paychecks. The cost of health insurance premiums is deductible to the payer, and the benefits received are tax-free, with certain exceptions for S Corporation Employees.By estimating the overall risk of health risk and health system expenses over the risk pool, an insurer can develop a routine finance structure, such as a monthly premium or payroll tax, to provide the money to pay for the health care benefits specified in the insurance agreement.The benefit is administered by a central organization, such as a government agency, private business, or not-for-profit entity.

A health insurance policy is a contract between an insurance provider (e.g. an insurance company or a government) and an individual or his/her sponsor (that is an employer or a community organization). The contract can be renewable (annually, monthly) or lifelong in the case of private insurance. It can also be mandatory for all citizens in the case of national plans. The type and amount of health care costs that will be covered by the health insurance provider are specified in writing, in a member contract or "Evidence of Coverage" booklet for private insurance, or in a national [health policy] for public insurance.

Choosing a health insurance plan can be tricky because of plan rules regarding in- and out-of-network services, deductibles, co-pays, and more.Since 2010, the Affordable Care Act has prohibited insurance companies from denying coverage to patients with pre-existing conditions and has allowed children to remain on their parents' insurance plan until they reached the age of 26.

2 A Brief 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.

Name of the Target Column: charges

Unit: **Float**

Description: Individual medical costs billed by health insurance.As value of money is float,that is why,the unit of the target column (Charges) is float.

3 Description of the Models used in this Project

Total 2 Regression models are used in this project. They are:

1. **Linear Regression**

The first model which is used in this project is Linear Regression Model.Linear Regression is a machine learning algorithm based on supervised learning.It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Mathematically, we can represent a linear regression as:

$$y = a_0 + a_1 x + \epsilon$$

Where,
y = Dependent Variable (Target Variable)
x = Independent Variable (predictor Variable)
a_0 = Intercept of the line (Gives an additional degree of freedom)
a_1 = Linear regression coefficient (scale factor to each input value)
ϵ = Random Error

2. **Random Forest Regression**

The second model which is used in this project is Random Forest Regression Model.Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea

of the bagging method is that a combination of learning models increases the overall result.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "***Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.***" Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.which means that Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

4 Performance Scores of Each Model

Regression Model	Mean Absolute Error (MAE)	Mean Squarred Error(MSE)	Root Meaned Squared Error(RMSE)	R^2 Score
Linear Regression	0.416854	0.348277	0.590	55.3%
Random Forest Regression	0.323846	0.261649	0.511	66.4%

5 Conclusion

Before drawing conclusion, we have to check the facts for four performance score.What kind of score makes a model better than other? For Mean Absolute Error, a perfect MAE value is 0.0, which means that all predictions matched the expected values exactly. For Mean Squared Error, a perfect MSE value is 0.0, which means that all predictions matched the expected values exactly. For Root Mean Squared Error, a perfect RMSE value is 0.0, which means that all predictions matched the expected values exactly. Lastly, the most common interpretation of r-squared is how well the regression model fits the observed data. For example, an r-squared of 64% reveals that 64% of the data fit the regression model. Generally, a higher R^2 (r-squared) indicates a better fit for the model.

So, the lower the score of MAE, MSE, RMSE is and the higher the score of R^2 is, the better the model will be.

Now,we can come to the conclusion that, with the given performance scores for the two (2) models, we can say that Random Forest Regression is the most suitable model for this dataset and Linear Regression is the least least suitable model for this dataset.