Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

CSE 4108

Artificial Intelligence Lab

Project Name: Medical Insurance Cost Prediction

Submitted To

Mr. Faisal Muhammad Shah
Associate Professor, CSE, AUST

Md. Siam Ansary

Lecturer, CSE, AUST

Submitted By

Nusrat Khan	170204101
Fuad Chowdhury	180104137
Abdur Samad	180104139

Date of Submission **:** 12/03/2022

Description of the Problem

Rising health care costs are a major economic and public health issue worldwide. Insurance is a policy that eliminates or decreases the costs of health care treatment occurred by various risks. Various factors can influence the cost of insurance. Machine learning for the insurance industry sector can make the prediction of insurance charges more efficiently. In this project, we have demonstrated different regression models to predict the insurance amount according to some features for our given dataset.

Dataset of the Problem

We have used Medical Insurance Cost Prediction Dataset to implement regression model. We have collected our data by doing a survey. The dataset has a total of 9 columns and 300 rows including the target column. The target column is the last column and rest of the columns are the features**.**

The target column is ‘Insurance Amount’ and the feature columns are:

- **Age** (Age of survey participants)
- **Children** (The number of children that person has)
- **BMI** (Body Mass Index of that Person)
- **Smoker** (Is that person a smoker or a non-smoker)
- **Diabetes** (Whether that person has diabetes or not)
- **Surgery** (Whether there has been any surgery history of that person)
- **Chronic Disease** (Whether the person has any chronic disease or not)
- **Gender** (Gender of that person)
- **Insurance Amount** (Individual medical costs billed by health insurance company)

Model Description

We used five regression models for predicting the insurance cost and they are:

1. Linear Regression
2. KNN Regression
3. Support Vector Machine
4. Decision Tree
5. Random Forest

Linear Regression: Linear Regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression

KNN Regression: KNN Regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

Support Vector Machine: Support Vector Machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier's detection. The advantages of support vector machines are: Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples.

Decision Tree: Decision Tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Random Forest: Random Forest is a supervised learning algorithm. It can be used for both classification and regression. It is also the most flexible and easy to use algorithm. Random forest classifier selects decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of feature importance.

Performance Comparison

	Linear Regression	KNN Regression	Random Forest	Decision tree Regression	Support vector Machine Regression
MSE	15484048.2384	22629901.7870	16974867.8093	22902608.7500	48007426.8235
RMSE	3934.9775	4757.0896	4120.0567	4785.6670	6928.7391
MAE	3116.7074	3471.9722	3189.1030	3388.2500	5910.9426
R square	0.68437587	0.53871604	0.65398727	.533157238	0.021425028

Discussion:

Here we successfully predict the Medical Insurance price using different machine learning algorithms. The project uses a total of 5 machine learning algorithms where each has different performance values. In the end the model's performance value are compared for better understanding of the models. After analysing the five models, we have come to a conclusion that the Linear Regression Model has performed better than the rest of the models. This model works pretty smooth and compute better results for datasets. The accuracy of the Linear Regression Model is 68% that means the model is predicting 68% of the data correct, which is an ideal value. After that, the accuracy of the Random Forest Regression is 65% that means the model is predicting 65% of the data correctly. Then, the accuracy of KNN Regression Model and Decision Tree Regression Model is almost same which 53%. The accuracy of Support Vector Regressor model is below than 10%. So, in compare to these 5 models for predictive dataset, we can say that Linear Regression Model is best suited than other for this dataset.

Contribution :

170204101 : (40%)
Dataset Collection, Data Pre Processing, Showing Co Relation, Applying ML algorithms (**Random Forest**), Providing visual differences in evaluation metrics (RMSE Score, R2 Score, MAE Error) for different ML Algorithms.

180104137 : (30%)
Dataset Collection, Applying ML algorithms (**KNN Regression:**, **Support Vector Machine**).

180104139: (30%)
Dataset Collection, Applying ML algorithms (**Linear Regression**, **Decision Tree**).