NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY,

BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA)

COURSE LA2 PROPOSAL

On

Medical Insurance Price Prediction

Submitted in partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

Anurag Nepal	1NT19CS036
Avishek Rijal	1NT19CS045
Baibhav Dhakal	1NT19CS048
Nabin Kumar KC	1NT19CS116

NITTE EDUCATION TRUST

Department of Computer Science and Engineering

2021-22

Nitte Meenakshi Institute of Technology

(AN AUTONOMOUS INSTITUTION AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY)

PB No. 6429, Yelahanka, Bangalore 560-064, Karnataka

Telephone: 080- 22167800, 22167860

Fax: 080 – 22167805

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Course Research Paper titled "Medical Insurance Price Prediction" is an authentic work carried out by Anurag Nepal(1NT19CS0363), Avishek Rijal(1NT19CS045),Baibhav Dhakal(1NT19CS048) And Nabin Kumar KC(1NT19CS116), Bonafide students of Nitte Meenakshi Institute of Technology, Bangalore in partial fulfilment for the award of the degree of Bachelor of Engineering in COMPUTER SCIENCE AND ENGINEERING of Visvesvaraya Technological University, Belagavi during the academic year 2021-22.

Name and Signature of the Faculty In charge **Name and Signature of the HOD**

Date:

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. We express our sincere gratitude to our Principal Dr. H. C. Nagaraj, Nitte Meenakshi Institute of Technology for providing facilities.

We thank our HoD, Dr. Sarojadevi H for the excellent environment created to further educational growth in our college. We also thank her for the invaluable guidance provided which has helped in the creation of a better technical report.

Finally, we thank our Subject Faculty, Mrs Vani V for providing us with the knowledge with which we were able to complete this proposal. We also thank all our friends, teaching and nonteaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the presentation.

Table of Contents

ABSTRACT

In this project, we analyse the personal health data to predict the medical insurance amount for individuals. Two regression models naming Multiple Linear Regression and Decision Tree Regression are to be used to compare and contrast the performance of these algorithms. Dataset is used for training the models and that training helps to come up with some predictions. Then the predicted amount is to be compared with the actual data to test and verify the model. Later the accuracies of these models are to be compared.

INTRODUCTION

The goal of this project is to allow a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance.

We are thinking of applying the data mining tools that we have been learning this semester to try and tackle this problem to predict the insurance cost of a person based on the age, sex, BMI, number of children ,whether he/she is a smoker and the region they preside in.

DATASET

The data set we chose from Kaggle has the columns as

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- BMI : 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
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

The dataset is not suited for the regression to take place directly. So cleaning of dataset becomes important for using the data under various regression algorithms.

METHODS AND MODELS

Regarding the model we are trying to find out the best model by comparing their output and their performance in both the train and test instances. We will use various variations of regression model and a decision tree classifier to find out the optimal algorithm.

Regression analysis allows us to quantify the relationship between outcome and associated variables. Many techniques for performing statistical predictions have been developed, but, in this project, two models – Multiple Linear Regression (MLR) and Decision tree regression are to be tested and compared.

Multiple linear regression can be defined as extended simple linear regression. It comes under usage when we want to predict a single output depending upon multiple input or we can say that the predicted value of a variable is based upon the value of two or more different variables. The predicted variable or the variable we want to predict is called the dependent variable and the variables being used in predict of the value of the dependent variable are called the independent variables.

Classification models in decision tree regression builds in the form of a tree structure. The dataset is divided or segmented into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision tree with decision nodes and leaf nodes is obtained as a final result. These decision nodes have two or more branches, each representing values for the attribute tested. Decision on the numerical target is represented by leaf node. The topmost decision node corresponds to the best predictor in the tree called root node. Numerical data along with categorical data can be handled by decision tress.

PRESENTATION AND VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose.

So we are thinking of representing the data and the results after the mining using charts, histograms to compare and contrast between the algorithms and their performance.

ROLES

Team Member	**Task**
Avishek Rijal	Multiple Regression
Anurag Nepal	Multiple Regression
Baibhav Dhakal	Decision Tree
Nabin Kumar KC	Decision Tree

SCHEDULE

Date	**Task to be Completed**
29/12/2021	Choose dataset and write proposal
10/01/2022	Data pre- processing to be completed
13/01/2022	Implement algorithms
15/01/2022	Completion of Project and the Report
17/01/2022	Presentation

REFERENCES

Project Proposal: https://www.fool.com/the-blueprint/project-proposal/

Dataset: https://www.kaggle.com/mirichoi0218/insurance

Decision Tree: https://www.saedsayad.com/decision_tree_reg.htm

Multiple Regression: https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

Visualization: https://www.tableau.com/learn/articles/data-visualization