**Decision Support System for Health Insurance**

UNIVERSITY OF NORTH TEXAS

CSCE 5215 (SECTION 004): MACHINE LEARNING

Professor (or Dr.): DR.SAYED KHUSHAL SHAH

**Team Members:**

1.Manisha Singam (11675497)

2.Lakshmi Prasanna Valdas (11700056)

3.Sekhar Reddy Kandula (11696582)

4.Abdullah Mohammed (11735588)

## Executive Summary

The primary problem the project aims to address is to help bring down the cost of health insurance for an individual. The solution for this problem is to build a Decision Support System which can predict the health insurance charges for a person and the contribution of various factors towards it using Machine Learning Models. The analysis and results of this work will help individuals and families make better healthcare decisions, improve their lifestyle, and bring down the average amount spent by an American on insurance.

## Project Background

Healthcare can be highly expensive in the US. An individual doctor's office visit could cost several hundred dollars. Since we don't know when we could get sick or how much care we might require, most of us would not be able to afford such high costs if we get sick. Such costs can be brought down to more manageable levels with the help of health insurance. However, there are many Health Insurance providers, and each considers several factors in determining the cost of insurance for an individual. U.S. health care spending grew 2.7 percent in 2021, reaching $4.3 trillion or $12,914 per person [1]. The analysis and results of this work will help individuals and families make better healthcare decisions, improve their lifestyle, and bring down the average amount spent by an American on insurance.

Hence, a Decision Support System for determining which factor contributes the most to Insurance costs would help an individual make better decisions.

**Problem Statement**- The primary problem addressed in this project is predicting the cost of health insurance for a person based on various factors like age, gender, body-mass index, number of children, smoking history, resident region. The goal of this work is to answer questions like-

1. Can we build a reliable decision support system for health insurance costs for individuals?
2. Which factor contributes most to determining health insurance amount for an individual?

On finding which factor contributes the most to Insurance costs an individual can make changes to his/her lifestyle, diet, location if possible and required and reduce their overall health insurance cost.

1. Which Machine Learning model helps predict more accurate health insurance costs?

**Input-Output pair-** The input for this project is factors that contribute to health insurance costs- 6 data features and 1338 data rows. The output is the charges and features ranked by importance of contribution to insurance charges.

A screenshot of a graph

Description automatically generated

Feature Importance-

A screenshot of a computer

Description automatically generated

Several attempts were made in this direction to build models to predict the insurance costs. In [4] trained and evaluated artificial intelligence network-based regression models are employed to predict health insurance premiums. In [5], the proposed model incorporates and demonstrates different models of regression such as Ridge Regression, Lasso Regression, Simple Linear Regression, Multiple Linear Regression and Polynomial Regression. For this model the RMSE value is 5100.53 and R -squared value is 0.80.

However, most of them fall short on multiple factors- The models are not accurate enough to rightly predict the charges, evaluation metrics used for evaluating the predictions involve a lot of bias, an ensemble model approach has been successfully tried before. Some works involved the usage of deep learning and neural network models which would increase the space and time complexity of the system [4].

***Baseline-*** Linear Regression is used as the baseline to compare the results against for predicting the health insurance costs.

*Baseline\_DSS for Health Insurance* file is attached along with this document.

## Project Solution

***Vision Statement-*** To build a decision support system which can help people make lifestyle changes and take better decisions which indeed help live and save better.

***Project Schedule-*** The major milestones for the project are- Data Collection and Preprocessing, Exploratory Data Analysis, Feature Selection, Model Selection, Model Training and Hyperparameter Tuning, Model Evaluation, Deployment and Recommendations.

file is included which contains the baseline implementation.

**Proposed Methods-** Ridge Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, DecisionTree Regressor, Gradient Boosting, Extreme Gradient Boosting and combination of models (Ensemble Model) will be used for this project.

**Evaluation Metric-** The performance will be evaluated usingMean Absolute Error,Mean Squared Error, Root-Mean-Square-Error**,** r2 score metrics against the baseline. Lower the score, better the model.

The results for evaluation metrics for the baseline model are-

**A screenshot of a computer error

Description automatically generated**

Mean Absolute Error, Mean Square Error and Root Mean Square Error can range from 0 to infinity. The lower the score, the better the prediction.

R2 score- can range between 0 and 1. The higher the score, the better the prediction.

****

Linear Regression score is 0.7613126015198816

**Data-** The dataset for this project is taken from the GitHub repository – Data for Machine Learning with R- <https://github.com/stedy/Machine-Learning-with-R-datasets>. The actual source of the dataset is from the book [Machine Learning with R](https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r) by Brett Lantz.

**Experiments**- Various quantitative experiments will be conducted. To name a few- Check for distribution of charges with respect to different features, Interval Estimation for the charges and Hypothesis testing will be performed. Quantitative variables will be generated for representing categorical data. Correlational analysis will be conducted as part of a non-experimental type of quantitative analysis.

**Qualitative Analysis-** Various qualitative experiments will be conducted. To name a few- Check for distribution of charges by applying logarithmic function, with respect to different features, histograms, barplots and several other plots will be constructed using Python libraires to analyze the quality of the data.

**Project Deliverables-** The outcome of the project is the prediction of insurance costs for an individual with given values for the various factors considered and finding the factors which contribute the most towards determining the insurance charges. On the basis of these results lifestyle changes can be suggested for an individual for better living and saving!

**Reporting Tools-** Jupyter Notebook is used for the project. Python language and its rich set of libraries are used to perform the analysis of data and build machine learning models.

**Team Roles and Responsibilities**

All group members plan to contribute equally. The individual contributions are-

Manisha Singam: Topic selection, perform data collection, co-author, and review the documents.

Lakshmi Prasanna Valdas: Topic selection, perform data collection, preprocessing, and data exploration.

Abdullah Mohammed: Topic selection, Perform data preprocessing and data exploration, co-author documents.

Sekhar Reddy Kandula: Topic selection, Perform Feature selection, model training, tuning, evaluation; co-author the documents.

**Project Goals and Deliverables**

1. **Data Collection and Preprocessing**

Utilize the data from - Data for Machine Learning with R- <https://github.com/stedy/Machine-Learning-with-R-datasets>.

A lot of cleaning will be done on the data to make it suitable for analysis and modelling- remove duplicates from the data, handle null data, one-hot encoding, scaling, convert object labels into categorical, convert categorical labels into numerical will be performed.

1. **Exploratory Data Analysis (EDA)**

EDA will be done to understand the data, visualize the data, distribution of charges, density plots, detect outliers, explore correlation between variables in the data.

1. **Feature Selection**

The greater the value of the coefficient, the higher the feature’s contribution towards the insurance costs. The features with higher positive correlation with the insurance costs will be selected, and those with negative correlation will be eliminated.

1. **Model Selection**

Machine Learning models will be employed on the data to find the correlation between the features and the insurance costs. Ridge Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, DecisionTree Regressor, Gradient Boosting, Extreme Gradient Boosting and combination of models (Ensemble Model).

1. **Model Training and Hyperparameter Tuning**

Train models using k-fold cross-validation. Regularize the model’s ability to structurally prevent overfitting by imposing a penalty on the coefficients. The models with optimal hyperparameter tuning will be evaluated by comparing the predictions of each model with validation data.

1. **Model Evaluation**

The models will be evaluated usingMean Absolute Error,Mean Squared Error, Root-Mean-Square-Error**,** r2 score metrics which describe the differences between the predicted values and the observed values.

1. **Deployment and Recommendations-**

After the data is analyzed in detail, the density, relations, and correlations between the featured are analyzed, the Machine Learning models will be built. Based on all the observations the decision support system will be built and deployed. Based on the DSS, several observations and recommendations can be made for health insurance costs.

**Timeline**

|  |  |
| --- | --- |
| Week 1 | Data Collection and Preprocessing |
| Week 2 | Exploratory Data Analysis |
| Week 3 | Feature Selection |
| Week 4 | Model Selection |
| Week 5 | Model Training and Hyperparameter Tuning |
| Week 6 | Model Evaluation |
| Week 7 | Deployment and Recommendations |
| Week 8 | Documentation |

## Conclusion

A Decision Support System for determining which factor contributes the most to Insurance costs will help an individual make better decisions. The decision support system for finding insurance costs and determining the greatest contributing feature towards insurance costs has been built in Python. On finding which factor contributes the most to Insurance costs an individual can make changes to his/her lifestyle, diet, location if possible and reduce their overall health insurance cost.

## References

CMS.gov. (n.d.). *National Health Expenditure Data*. Historical. <https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/historical#:~:text=U.S.%20health%20care%20spending%20grew,spending%20accounted%20for%2018.3%20percent>.

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Roger, Z., Pavlov & Nacho, C. (2017, March 28). *Machine-Learning-with-R-datasets*. GitHub. <https://github.com/stedy/Machine-Learning-with-R-datasets>

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health*, *19*(13), 7898.

Panda, S., Purkayastha, B., Das, D., Chakraborty, M., & Biswas, S. K. (2022, May). Health insurance cost prediction using regression models. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)* (Vol. 1, pp. 168-173). IEEE.