# **Decision Support System for Health Insurance**

**Team Members:**

1.Manisha Singam (11675497)

2.Lakshmi Prasanna Valdas (11700056)

3. Sekhar Reddy Kandula (11696582)

4. Abdullah Mohammed (11735589)

UNIVERSITY OF NORTH TEXAS

CSCE 5215(Section 004): Machine Learning

Professor (or Dr.): Dr. Sayed Khushal Shah

## Motivation

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.

## Significance

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](#_References)]. 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.

**Objectives**- The primary objective of 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?

**Features-** The features of the project are- 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!

**Related Work-** Several attempts were made in this direction to build models to predict the insurance costs. In [[4](#_References)] trained and evaluated artificial intelligence network-based regression models are employed to predict health insurance premiums. In [[5](#_References)], 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 [[2](#_References)].

**Dataset-** 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 [[3](#_References)].

*Input-Output pair-*The input for this project is factors that contribute to health insurance costs- 13 data features and 15000 data rows. The output is the charges and features ranked by importance of contribution to insurance charges.

A screenshot of a computer

Description automatically generated

Feature Importance-

A screenshot of a computer code

Description automatically generated

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.

1. **Data Collection and Preprocessing**

Utilize the data from - Data for Machine Learning with R- <https://github.com/stedy/Machine-Learning-with-R-datasets>. [[3](#_References)]

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.

***Importing the data-***

***A screenshot of a computer

Description automatically generated***

***Understanding the data-***

A screenshot of a computer

Description automatically generated

***Removing duplicates from the dataset-***

A screenshot of a computer code

Description automatically generated

***Detect outliers in the dataset-***

A screenshot of a computer

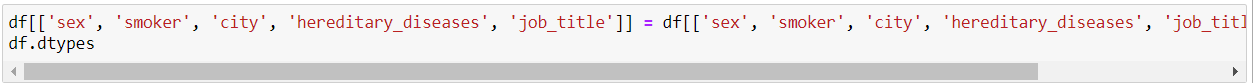
Description automatically generated

***Remove outliers in the dataset-***

A screenshot of a computer

Description automatically generated

***Converting objects labels into categorical-***



A screenshot of a computer code

Description automatically generated

***Converting category labels into numerical using LabelEncoder-***

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a computer code

Description automatically generated**

***Handling null values in the data-***

A close-up of a computer code

Description automatically generated

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.

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

***Qualitative Analysis-*** Various qualitative experiments are 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 are constructed using Python libraires to analyze the quality of the data.

***Statistical Distribution-*** The distinct value shows number of distinct values in a attribute Skewness provide the Mode of the attribute by comparing with Mean we can conclude if its right or left skewed and explains if the distribution is peak or flat related to normal distribution

**A screenshot of a computer screen

Description automatically generated**

**A table with numbers and letters

Description automatically generated**

***Histogram for distinct-***

***A graph with numbers and a bar

Description automatically generated***

***Histogram for skewness-***

***A graph with blue bars

Description automatically generated***

***Histogram for kurtosis-***

***A graph with blue bars

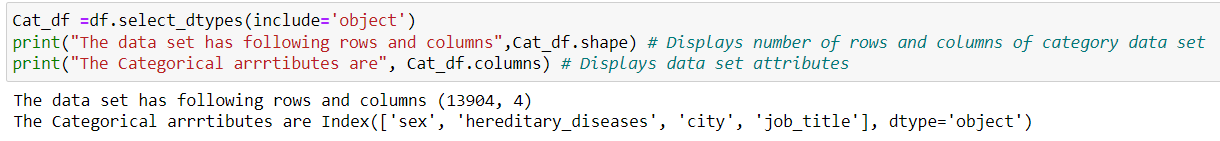
Description automatically generated with medium confidence***

***Histogram for mean-***

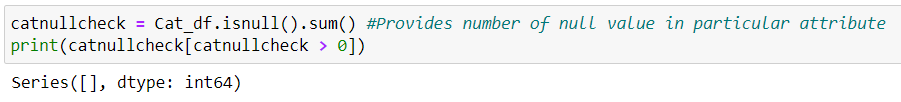
***A graph with numbers and a bar

Description automatically generated***

***Categorical Data Analysis-***

******

***Null value analysis-***

******

***Numerical Data Analysis-***

***A screenshot of a computer code

Description automatically generated***

***Null value analysis-***

***A screenshot of a computer

Description automatically generated***

***Null percentage check-***

***A screenshot of a computer code

Description automatically generated***

***Distribution of charges-***

***A graph with blue lines

Description automatically generated***

***Distribution of charges with log applied-***

***A graph with red lines

Description automatically generated***

***Distribution for age using density plot-***

***A graph with a line

Description automatically generated***

***Distribution for bmi using density plot-***

***A graph with a line

Description automatically generated***

***Distribution of charges for patients with BMI greater than 30***

***A graph with a line graph

Description automatically generated***

***Distribution of charges for patients with BMI less than 30-***

***A graph showing the amount of charge

Description automatically generated***

***Distribution of charges by city-***

***A graph of blue and white bars

Description automatically generated with medium confidence***

***Barplot for distribution of charges by city and sex-***

***A graph with text on it

Description automatically generated***

***Barplot for distribution of charges by city and smoker-***

***A graph of different colored lines

Description automatically generated***

***Barplot for distribution of distribution of charges by city and children-***

***A graph with black lines

Description automatically generated with medium confidence***

***Analyzing the medical charges by age, bmi and children according to the smoking factor-***

***A graph with red and blue dots

Description automatically generated***

***A graph with green and orange dots

Description automatically generated***

***A graph of smoke and children

Description automatically generated***

***Violin plot for analyzing charges with children and smoking habit-***

***A graph of different colored lines

Description automatically generated***

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.

***Finding correlation between variables-***

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

***Feature Importance Ranking-***

***A screenshot of a computer

Description automatically generated***

A graph with red bars

Description automatically generated

1. **Model Selection**

Linear Regression, Lasso Regression, Ridge Regression, ElasticNet, KernelRidge, Support Vector Regression, Gradient Boosting, Extreme Gradient Boosting, Random Forest Regression, 4 types of Naïve Bayes- GaussianNB, MultinomialNB, BernoulliNB, CategoricalNB, Stochastic Gradient Descent, Perceptron and combination of models (Ensemble Model) will be used for this project.

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

***Lasso Regression-*** adds the penalty equivalent to the absolute value of the sum of coefficients. Uses L1 regularization.

***Ridge Regression-*** adds the penalty equivalent to the linear least squares function. Uses L2 regularization.

***Elastic Net-*** is the combination of both Ridge and Lasso Regressions. It adds both the sum of squared coefficients and the absolute sum of the coefficients with the ordinary least square function. It is useful when there are multiple features that are correlated.

***Kernel Ridge Regression-*** is a non-parametric form of ridge regression. The aim is to learn a function in the space induced by the respective kernel k by minimizing a squared loss with a squared norm regularization term.

***Support Vector Regression*-** finds a hyperplane that minimizes prediction error and approximates the relationship between the input variables and a continuous target variable.

***Gradient Boosting-*** is generally used when we want to decrease the bias error. In each stage, a regression tree is fit on the negative gradient of given loss function.

***Extreme Gradient Boosting-*** is a scalable tree boosting algorithm. It is designed to be both computationally efficient and highly effective.

***Light Gradient Boosting Machine-*** is based on decision tree algorithms and used for ranking, classification, and other machine learning tasks.

***Random Forest Regression-* contains many individual decision trees where each tree spits out a class prediction and the class with the most votes becomes our model’s prediction.**

***Naive Bayes-* is a technique based on the Bayes theorem.** 4 types of Naïve Bayes- ***GaussianNB, MultinomialNB, BernoulliNB, CategoricalNB***

***Stochastic Gradient Descent-*** the training set is run one data point at a time and the parameters are updated according to the error gradient.

***Perceptron-*** is a single layer neural network. It learns optimal weight coefficients automatically, multiplies the weights with input features and decides.

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.

**A white screen with black text

Description automatically generated with medium confidence**

Training models using k-fold cross-validation and seeing the performance of all the models in the given dataset.

**A white rectangular sign with black text

Description automatically generated**

**A screen shot of a computer code

Description automatically generated**

1. **Model Evaluation**

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

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

A screenshot of a computer program

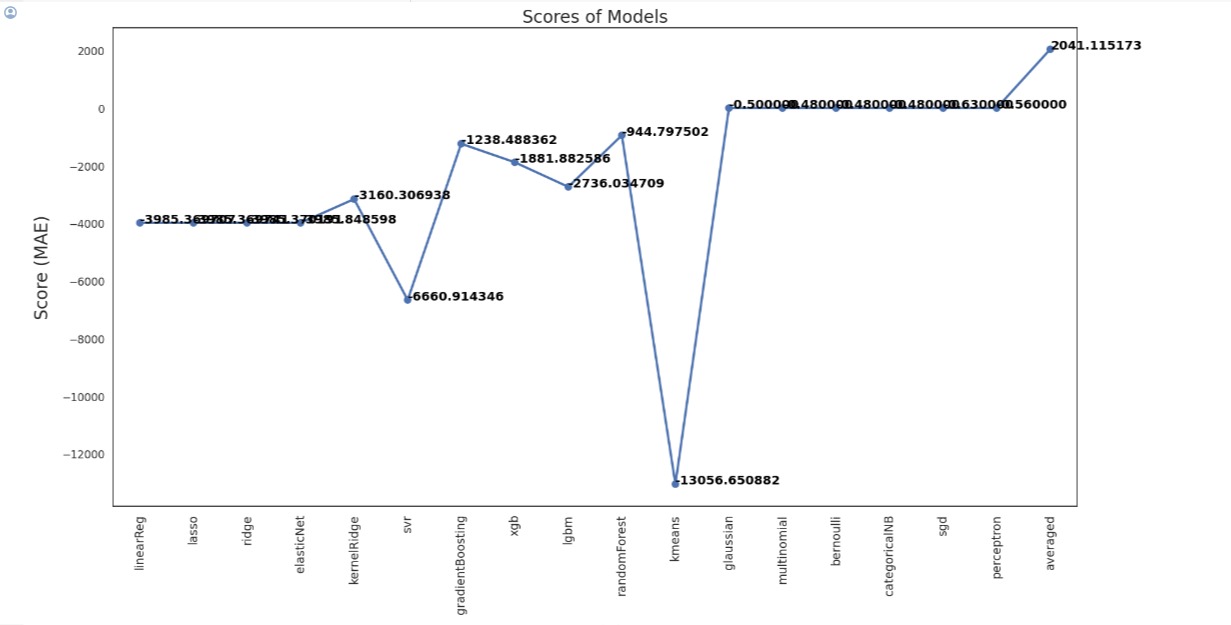
Description automatically generated

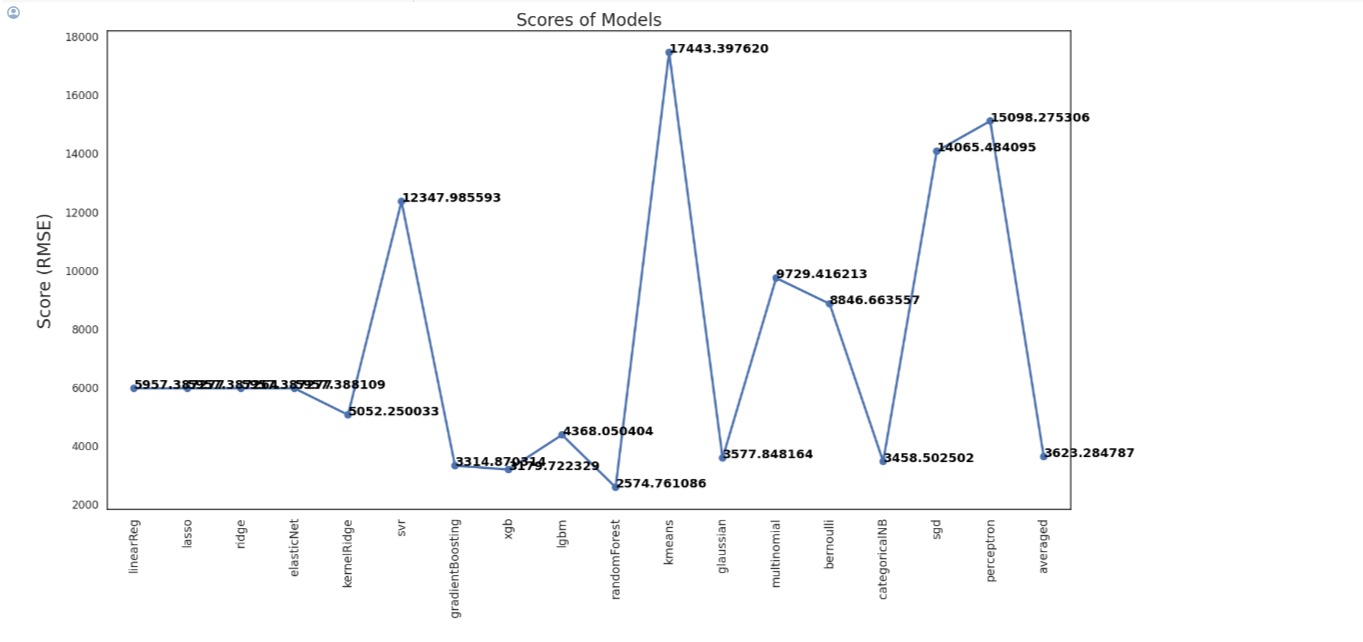
Mean Absolute Error, Mean Square Error and Root Mean Square Error can range from 0 to infinity. The lower the score, the better the prediction.

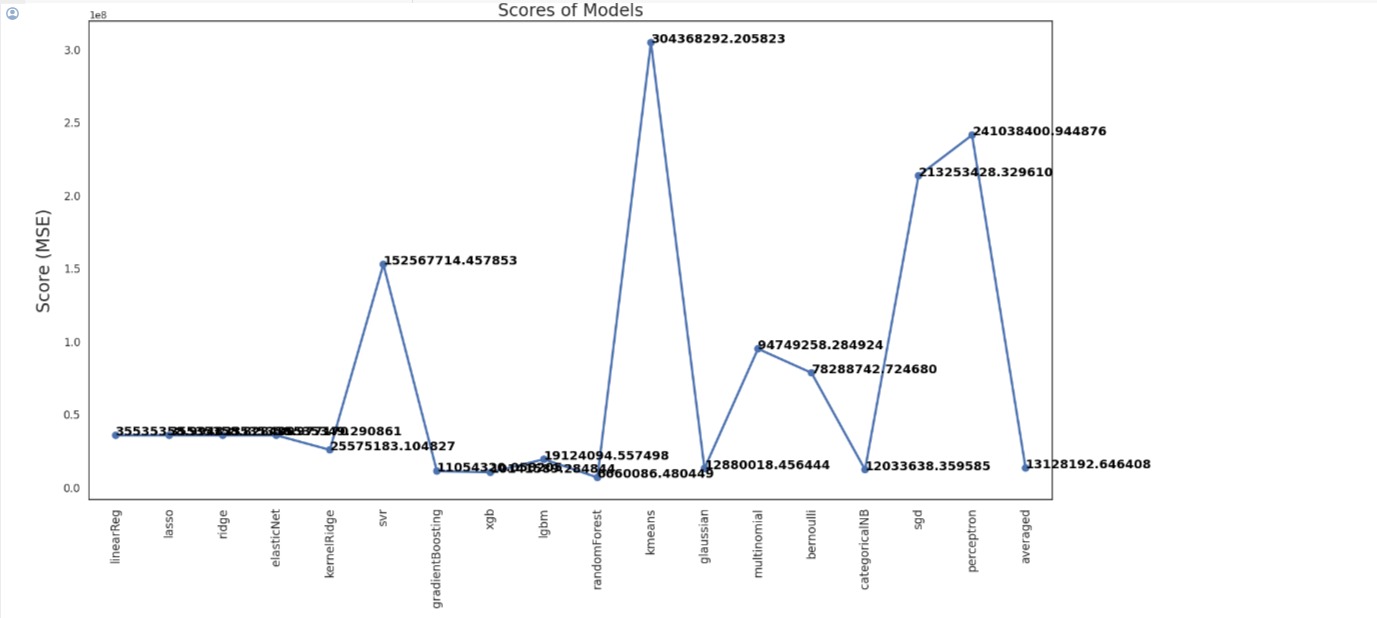
The results for evaluation metrics for the baseline model are-

A white screen with black text

Description automatically generated







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-

* Distribution of charges in the dataset is right-skewed.
* People with children tend to have higher insurance charges.
* Smoking has the greatest influence on insurance. Costs also increase with Age, bmi and number of children.
* From the violin plot we can observe that people who have children tend to smoke less.
* From the Correlation Matrix diagram, we observe that Smoking has the highest correlation with insurance costs.
* The order of feature influence based on Linear regression fit is smoking, children, bmi and age respectively.
* The order of feature influence based on Lasso regression fit is smoking, children, bmi and age respectively.

Based on the DSS, the relations between features, correlation matrix, and regression models all show that ‘Smoker’ followed by ‘bmi’, followed by ‘age’ followed by ‘children’ are the greatest contributors to insurance costs for an individual.

The decision support system for 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 required and reduce their overall health insurance cost.

The results from the decision support system show that ***Smoking*** is the most contributing feature for insurance costs. Hence, when individuals reduce their ***smoking*** habit, they will not only have better health but also reduce their insurance costs.

**Project Management**

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

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.

***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.

**Work Completed-**

***Description-*** The Model Training and Hyperparameter Tuning, Model Evaluation must be completed for the remaining models and Deployment and Recommendations must be done.

***Responsibility-*** All group members plan to distribute the remaining tasks equally. The individual responsibilities would be-

**Manisha Singam:** Perform model training, tuning, evaluation for combination of models (Ensemble Model); Perform final deployment and recommendations module; co-author the document.

**Lakshmi Prasanna Valdas:** Perform model training, tuning, evaluation for combination of models (Ensemble Model); Perform final deployment and recommendations module; co-author the document.

**Sekhar Reddy Kandula:** Perform model training, tuning, evaluation for combination of models (Ensemble Model); Perform final deployment and recommendations module; co-author the document.

**Abdullah Mohammed:** Perform model training, tuning, evaluation for combination of models (Ensemble Model); Perform final deployment and recommendations module; co-author the document.

***Contributions-*** All group members have contributed equally for increment II of the project, total – 30% of work.

**Manisha Singam –** 7.5%

**Lakshmi Prasanna Valdas –** 7.5%

**Sekhar Reddy Kandula –** 7.5%

**Abdullah Mohammed –** 7.5%

***Issues and concerns-*** We do not foresee any issues or concerns at this stage of the project.

## 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. <http://www.bdax.com.au/library/ebooks/Machine%20Learning%20with%20R.pdf>

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. <https://www.mdpi.com/1660-4601/19/13/7898>

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. <https://www.researchgate.net/publication/362707157_Health_Insurance_Cost_Prediction_Using_Regression_Models>