Health Insurance Cost Prediction using Machine Learning and Regression Models

Project Report

In Partial Fulfillment of the Requirements

For the Course

**STATISTICAL LEARNING (COMP-8590)**

For the Degree of

**MASTER OF APPLIED COMPUTING (MAC)**

**COMPUTER SCIENCE DEPARTMENT**

Submitted By

**MUZAKKIR QADRI MOHAMMED (110059185)**



**TABLE OF CONTENTS**

1. Abstract

2. Introduction

3. Problem Statement:

3.1. Problem Definition/Description/Formulation

3.2. Motivations

3.3. Justifications

4. Literature Survey

5. Methodology:

5.1. Material and Data

5.2. Proposed Methods/Solutions/Algorithms/Models

5.3. Conditions and Assumptions

6. Computational Experiments

6.1. Experiments

6.2. Evaluation Metrics

6.3. Implementation Details

6.4. Results

7. Conclusion:

7.1. Summary

7.2. Future Research

7.3. Open Problems

8. References

**1. ABSTRACT**

Health insuranceis a type of insurance coverage that pays for medical and surgical expenses incurred by the insured. Implementing Machine Learning for Health Insurance estimation can make the Insurance companies have more robust and well-made policies. This Project discusses the various approaches of Supervised Machine Learning in forecasting the Health Insurance costs. It demonstrates how different models of regression can forecast insurance costs and we will compare the results of models. The Models used in this project are Multiple Linear Regression, Decision Tree Regressor, Support Vector Machine, Random Forest Regressor, Random Forest Regressor using Principle Component Analysis and Stochastic Gradient Boosting.The best Technique observed during experimentation was Gradient Boosting Regressor with R- squared value of 0.903685, RMSE value of 0.321672, MSE value of 0.103473, MAE value of 0.122002.

**2. INTRODUCTION**

We live on a planet where health can never be guaranteed due to various unavoidable factors/ risks which may include any physical injuries, accidents, diseases, loss of property, loss of life, loss of health etc. There might be situations where these risks cannot be mitigated, so the World of Finance has come up with various products and policies to safeguard human from disaster while reducing the costs incurred. This is done by using Financial Capital to reimburse them for their loss. Therefore insurance can be defined as a policy in which an individual or an entity receives financial protection/aid/ reimbursement against risks and losses. This insurance capital is handled by an Insurance company and all decisions regarding policies are done by these Insurance companies. These Insurance policies act as barrier against the risk of financial loss, both big and small that may arise from damage to property injuries, damage to health etc. These Insurance companies are categorized based on the risk factors and each of these company provide insurance policies to individuals related to their agenda. One such Insurance company is the Health Insurance Company which provide health care insurance which mitigates the losses incurred due to health issues including any serious illness, injuries caused by accidents etc. All Insurance companies including the Health Insurance companies are responsible for quantifying the insurance cost estimates as per the policy. This quantification of cost estimates can be quantified and automated by using Machine Learning. Machine Learning can reduce human effort while increasing the credibility of the Health Insurance Companies. For any ML model to be trained, a suitable amount of relevant data is very essential. The data must have various Independent variables and a dependent variable (also called as target variable). Preprocessing is done on this data and then given as input to various regression models for training. The regression models will analyze the Independent variables and their corresponding target variables for patterns (also known as regression function). A part of the original data is kept aside as test data which will be used to evaluate the regression models. The evaluation metrics used are Root Mean Square Error, Mean Square Error, Mean Absolute Error and R-Squared Score. This will be further explained in detail in further sections of this report.

**3. PROBLEM STATEMENT**

**3.1. Problem Definition:**

The Quantification of insurance estimates can be a tedious task for any Insurance company. It requires experienced personnel for estimating insurance costs. These experienced personnel must consider each and every factor during their computation. If any factor is omitted when the amounts are calculated, the policy changes overall and can lead to fraud or even loss of life. It is therefore critical that these tasks are performed with high accuracy but humans are always bound to make errors.

**3.2. Motivations:**

Since decades, Health insurance companies have been relying on the traditional methods and actuarial formulas to estimate life expectancy and devise underwriting rules for their health insurance cost estimations. However, the conventional techniques are time-consuming, usually taking over a month and also costly. Hence, it is essential to find ways to make the underwriting process faster and more economical. Humans are always bound to make errors. If the task is provided to a trained machine then it is proven that a machine will outperform the human. Hence if the health care insurance cost estimation is automated then it will help the industry to focus on building better policies backed up by significant evidence obtained from Machine Learning. As the insurance industry continues to grow globally, it faces challenges related to fraud, high costs of sales, and pricing.

**3.3. Justifications:**

Given the complexities of the insurance industry, advanced analytics including Machine Learning algorithms, offers solutions to the challenges. Machine Learning is beneficial as it can automate the process of insurance estimations, thereby reducing human resources and error caused by humans. Various models will be made and compared with each other to obtain the model with the best performance. This will undoubtedly be of significance to Insurance companies .

**4. LITERATURE SURVEY**

Several papers have discussed the issue of claim prediction.

**4.1. Predict Health Insurance Cost by using Machine Learning and DNN Regression Models**

Link: https://www.researchgate.net/publication/348559741

Authors: Mohamed Hanafy, Omar M. A. Mahmoud.

The research uses various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set from Kaggle.com. The findings are summarized in Table IV. Shows that Stochastic Gradient Boosting offers the best efficiency, with an RMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.82. Stochastic gradient boosting can therefore be used in the estimation of insurance costs with better performance than other regression models. Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. Machine learning can significantly minimize these individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

**4.2. Health Insurance Amount Prediction**

Link: https://www.ijert.org/health-insurance-amount-prediction

Authors: Nidhi Bhardwaj, Rishabh Anand.

In this project, three regression models are evaluated for individual health insurance data. The health insurance data was used to develop the three regression models, and the predicted premiums from these models were compared with actual premiums to compare the accuracies of these models. It has been found that Gradient Boosting Regression model which is built upon decision tree is the best performing model. Various factors were used and their effect on predicted amount was examined. It was observed that a person’s age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features. The effect of various independent variables on the premium amount was also checked. The attributes also in combination were checked for better accuracy results. Premium amount prediction focuses on persons own health rather than other company’s insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount.

**4.3. Predicting Health Care Costs using Evidence Regression**

Link: <https://link.springer.com/article/10.1007/s40747-018-0072-1>

Authors: Belisario Panay, Nelson Baloian Jose A. Pino, Sergio Panafiel, Horacio Sanson, Nicola Bersano .

This research has specific implications for the business environment. Data analytics is now the trend that is gaining significance among companies worldwide. In the life insurance domain, predictive modeling using learning algorithms can provide the notable difference in the way which business is done as compared to the traditional methods. Previously, risk assessment for life underwriting was conducted using complex actuarial formulas and usually was a very lengthy process. Now, with data analytical solutions, the work can be done faster and with better results. Therefore, it would enhance the business by allowing faster service to customer, thereby increasing satisfaction and loyalty. The supervised learning algorithms namely, Multiple Linear Regression, Artificial Neural Network, REPTree, and Random Tree were implemented. The model validation was performed using tenfold cross-validation. The performance of the models was evaluated using MAE and RMSE. Findings suggested that the REPTree algorithm had the highest accuracy with lowest MAE and RMSE statistics of 1.5285 and 2.027, respectively, for the CFS method. Conversely, for the PCA method, Multiple Linear Regression showed the best performance with MAE and RMSE values of 1.6396 and 2.0659, respectively. Ultimately, it can be concluded that machine learning algorithms can be efficient in predicting the risk level of insurance applicants.

**4.4.** **Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression**

Link: https://www.mdpi.com/2227-9091/7/2/70/pdf

Authors: Jessica Pesantez-Narvaez, Montserrat Guillen, Manuela Alcañiz

XGBoost, and other boosting models, are dominant methods today among machine-learning algorithms and are widely used because of their reputation for providing accurate predictions. This novel algorithm is capable of building an ensemble model characterized by an efficient learning method that seems to outperform other boosting-based predictive algorithms. Unlike the majority of machine learning methods, XGBoost is able to compute coefficient estimates under certain circumstances and, therefore, the magnitude of the effects can be studied. The method allows the analyst to measure not only the final prediction, but also the effect of the covariates on a target variable at each iteration of the boosting process, which is something that traditional econometric models (e.g., generalized linear models) do in one single estimation step. When a logistic regression and XGBoost compete to predict the occurrence of accident claims without model-tuning procedures, the predictive performance of the XGBoost (tree booster) was much higher than the logistic regression in the training sample, but considerably poorer in the testing sample. Thus, a simple regularization analysis has been proposed here to correct this problem of overfitting. However, the improvement in predictive performance of the XGBoost following this regularization was similar to that obtained by the logistic regression. This means additional efforts have to be taken to tune the XGBoost model to obtain a higher predictive performance without overfitting the data. This might be considered as the trade-off between obtaining a better performance, and the simplicity it provides for interpreting the effect of the covariates

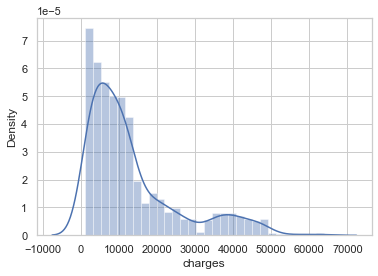
**5. METHODOLOGY**

**5.1. Material and Data:**

To create the claim cost model predictor, we obtained the data set through the Kaggle site. The data set includes seven attributes. The dataset is split into two parts i.e., the training set and the test set. The split is done in a way that 75% of the dataset is used for training the models and 25% is used for testing them. The training set is used to build the regression models so that they can predict the Health insurance cost and the test set is used to evaluate the regression model. The following table shows the Description of the Dataset It has the following predictors:

|  |  |
| --- | --- |
| Predictors | Description |
| Age | Age of the Sample client. |
| BMI | Body Mass Index |
| Kids | Number of Kids |
| Gender | Male or Female |
| Smoker | Yes or No |
| Region | Southeast, Southwest, Northeast, Northwest |
| Charges | The Insurance Cost |

The following figure shows the density variance of the dependant variable Charges.



**5.2. Proposed Methods/Solutions/Algorithms/Models**

The following 6 supervised Learning Techniques have been adopted:

**Multiple Linear Regression:**

The process of understanding if the independent variables (Age, BMI, Kids, Gender, Smoker and Region), (linearly) are related to the dependent variable (charges). This is referred to as the multiple linear regression (MLR) model. An MLR model with t independent features,,.., and Y results can be calculated as in the following equation.



In the above equation, u is the residual regression while a is the weight of each independent variable or parameter assigned

**Support Vector Regressor:**

SVMs can be generalized to problems with regression (i.e., when the outcome is continuous such as our target variable in our dataset). Essentially, SVMs seek hyperplanes i.e., they transform input dataspace into an extended function space to find out the hyperplane of separation. This results in better segregation of the input dataset. Specific functions called kernel functions are used to build this extended, separated functionality.

**Decision Tree Regressor:**

Decision Trees are straightforward, very popular, fast-training, and easy to read models with comparative or other methods of learning from the data. Decision Trees split the dataset into a Tree with each internal node being a Condition and each external node is the outcome. They are fairly competent but vulnerable to overfitting in their predictions.

**Random Forest Regressor:**

Random forests is a bagged decision trees that create a broad number of de-correlated trees so predictive performance is boosted further in comparison to decision trees.They are a very popular algorithm. They perform with good predictive performance and relatively few hyper-parameters. Several implementations of random forests can be found. Random forests take the average of the predictions obtained by multiple individual decision trees regressors thereby increasing the performance. As in the following equation , a random model for forest regressors can be expressed.



Where g is the final model made up of all the decision trees.

**Gradient Boosting Regressor:**

Gradient boosting machines are extremely well known ML regression model, Gradient boosting Regressor constructs a set of shallow trees. Each tree learns from its predecessor tree and further develops. While shallow trees are lesser predictive models, they can be "boosted" to create a stronger model, which if properly tuned , is difficult to tackle with other algorithms. The training of algorithms on a random training subsample provided more reductions in the tree correlation and thus enhanced predictive accuracy. This procedure is called a stochastic gradient boosting.

**Random Forest Regressor using PCA:**

PCA works by calculating eigenvalues of the correlation matrix of the attributes. The variance explained by each newly generated component is determined and the components retained are those which describe the maximal variation in the data set. PCA method is useful when used with predictive algorithms. Random Forest is performed on the features obtained after performing PCA.

**5.3. Conditions and Assumptions**

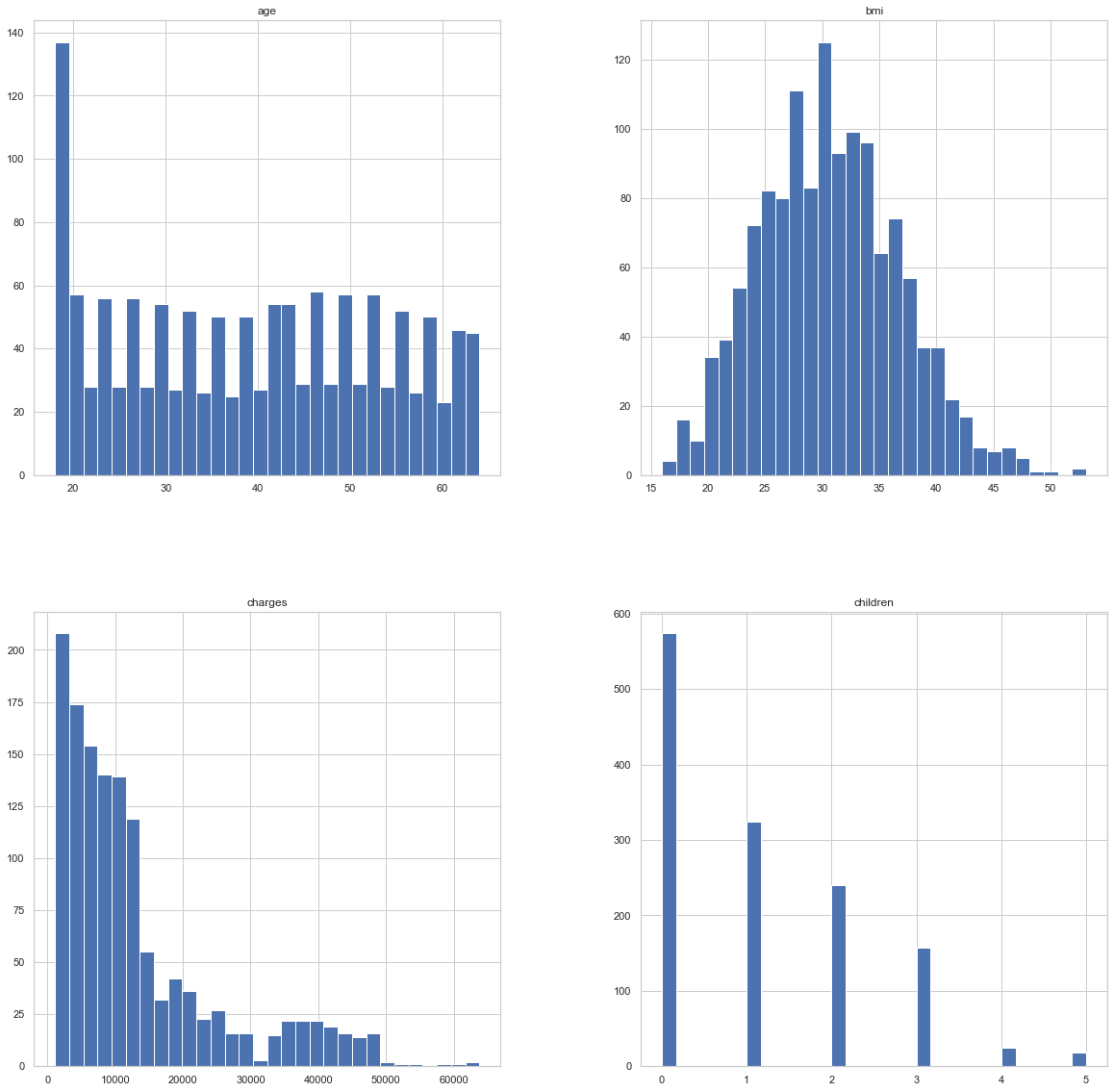
Assumption that the Predictors provided in the Dataset are the only Major Factors that affect the Insurance Estimation.There might be other real time factors that affect the Health thereby affecting Insurance estimations other than the ones specified in the dataset.

**6. COMPUTATIONAL EXPERIMENTS**

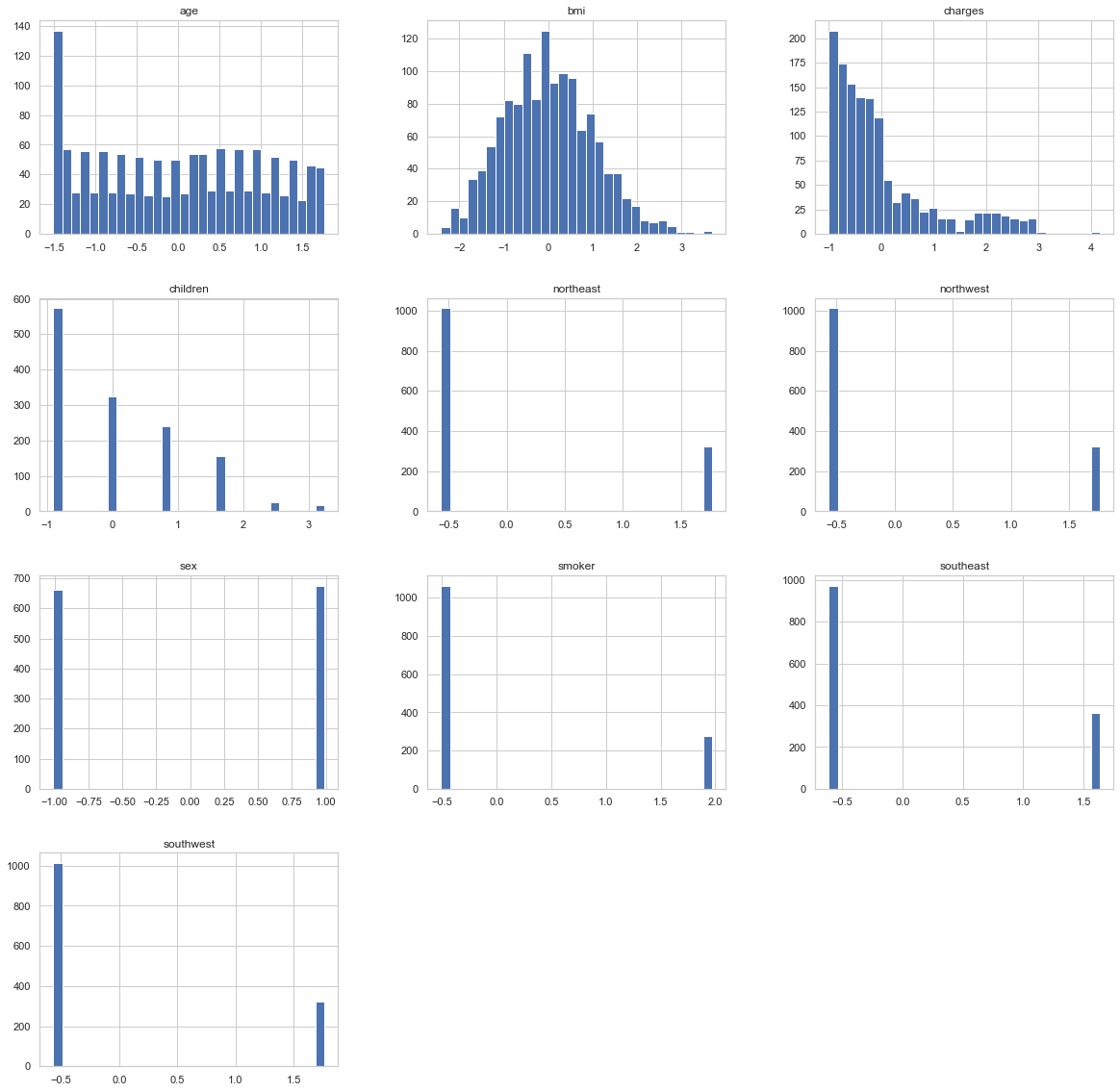
**6.1. Experiments:**

**Data Preprocessing:**

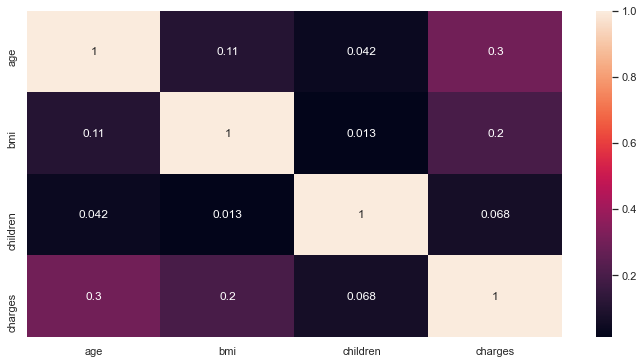
Before feeding the dataset to the models for training, they must be properly manipulated for any missing values and must be standardized to values relative to the dataset. Standardization is the process of rescaling the data items such that the mean of the data items becomes 0 with a standard deviation of 1. Before Standardization is performed, we perform feature expansion for the independent variable region of the dataset. The below figure shows the variance of the data before standardization.



The Below data shows the variance of the data items after standardization has been performed.



Correlation between the different predictors can be seen in the below figure.



As can be seem from the above figure, the predictors that are highly related to the target variable Charges are Age and BMI.

**Model Creation and Training:**

The Six models are created and the data is split into training set and test set with a split of 75:25 .The Training Data is given as input to the models for training them.

**Model Testing:**

The models are tested using the test sets . The graphs of predicted vs actual values for each of the six models can be seen below.

|  |  |
| --- | --- |
| Gradiant Boosting Regressor | Support Vector Regressor |
|  |  |
| Random Forest Regressor | **Random Forest with PCA** |
|  | C:\Users\qadri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\randomforest.png |
| Decision Tree | **Multiple Linear Regression** |
| C:\Users\qadri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\tree.png | C:\Users\qadri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\linear.png |

**6.2. Evaluation Metrics:**

The following metrics have been used to evaluate the performance of various models:

**R- Squared Score**: Measure of variance for a dependent variable that is explained by the predictor. The higher the value the better the model fits the data.

= 1 -

**Mean Squared Error**: Square of Difference between predicted value and actual value. The lower the value the better the model performs.

MSE =

n = Total Number of Data Points

= Observed Values

= Predicted Values

**Root Mean Squared Error:** It is the Square root of Mean Square Error. The lower the value the better the model performs.

RMSE =

RMSE=

**Mean Absolute Error**: It is the arithmetic average of the absolute errors. The lower the value the better the model performs.

MAE =

Values

= True Values

n= Total number of Data Points

**6.3. Implementation details:**

The Platform used is Windows 32 bit Operating System.

The Software used for development is Python Anaconda Spyder IDE.

The libraries used for model implementation are:

* scikit-learn
* numpy
* pandas
* The libraries used for Visualization are:
* seaborn
* matplotlib

**6.4. Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Model | MSE | RMSE | MAE | R-Square |
| Gradient Boosting Regressor | 0.103473 | 0.321672 | 0.122002 | 0.903685 |
| Support Vector Regressor | 0.121008 | 0.347863 | 0.199852 | 0.887362 |
| Random Forest Regressor | 0.122114 | 0.349448 | 0.236742 | 0.886333 |
| Random Forest Regressor using PCA | 0.164342 | 0.405391 | 0.281823 | 0.847027 |
| Decision Tree Regressor | 0.198318 | 0.445329 | 0.289724 | 0.815401 |
| Linear Multiple Regression | 0.219292 | 0.468286 | 0.331479 | 0.795878 |

As can be seen from the above Results, The highest performance can be observed by the Gradient Boosting Regressor, followed by Support Vector Regressor. The performance difference between Support Vector Regressor and Random Forest Regressor is negligible.

**7. CONCLUSION**

**7.1. Summary:**

Hence by performing various supervised techniques, we have obtained various Machine Learning Models that predict the insurance estimates over the health insurance Dataset. The best Technique observed during experimentation was Gradient Boosting Regressor with R- squared value of 0.903685, RMSE value of 0.321672, MSE value of 0.103473, MAE value of 0.122002. Therefore Gradient boosting regressor can therefore be used to estimate the insurance costs with better performance than other regression models.Using Machine Learning to predict the insurance costs can cut times taken in formulating plans for each individual.. Machine learning can significantly reduce human intervention for estimating costs of insurance and thereby help them to focus mainly on robust policy making based on the results from the prediction model. ML models thereby reduce time and resources. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

**7.2. Future Research:**

The Employment of more advanced supervised learning methods such as Deep Learning can further this project by giving it further insight.Addition of More Features that affect the insurance estimation can be done so that the models perform better.Can be futher evolved into Fraud Insurance Detection models.

**7.3. Open Problems :**

The main problem is that only one data set is being used. Data can be obtained from various sources like hospitals/ Insurance companies to further increase the size of the dataset. More Experiments on various other datasets can provide us more solid evidence on the comparative performance of different algorithms for predicting the Health care Insurance .

**8. REFERENCES**

* Kaggle Medical Cost Personal Datasets. Kaggle Inc. <https://www.kaggle.com/mirichoi0218/insurance>.
* Mohamed Hanafy and Omar M. A. Mahmoud(January 2021) Predict Health Insurance Cost by using Machine Learning and DNN Regression Models (IJITEE)
* Nidhi Bhardwaj and Rishabh Anand (June 2020) Health Insurance Amount Prediction (IJERT)
* Belisario Panay, Nelson Baloian Jose A. Pino, Sergio Panafiel, Horacio Sanson and Nicola Bersano (November 2019) Predicting Health Care Costs using Evidence Regression (MDPI)
* Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE
* Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.
* Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140
* Breiman, Leo, and others. 2001. ―Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).‖ Statistical Science 16 (3). Institute of Mathematical Statistics: 199–231.
* Friedman. 2002. ―Stochastic Gradient Boosting.‖ Computational Statistics & Data Analysis 38 (4). Elsevier: 367–78.
* Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. International Journal of Advanced Computer Science and Applications, 9(2).
* Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press.
* Shi X, Guo Z, Nie F, Yang L, You J, Tao D (2016) Twodimensional whitening reconstruction for enhancing robustness of Principal Component Analysis. IEEE Trans Pattern Anal Mach Intell 38:2130–2136 38.
* Yi S, Lai Z, He Z, Cheung Y, Liu Y (2017) Joint sparse principal component analysis. Pattern Recogn 61:524–536