# Utilizing Random Forest Regression for Health Insurance Premium Projections

Mrs.Breethy.S.V breethy@stellamaryscoe.edu.in
J. Sunanthini SUNANTHINI@stellamaryscoe.edu.in
Mr.R.Dinesh dinesh@stellamaryscoe.edu.in
Dr. F. R. Shiny Malar SHINYMALAR@stellamaryscoe.edu.in
Mrs.R.Shiny shiny@stellamaryscoe.edu.in

**Department Of Computer Science Engineering**
**Stella Mary's College Of Engineering, Tamilnadu, India**

Abstract: Artificial intelligence (AI) and machine learning (ML) in healthcare are approaches to make people's lives easier by anticipating and diagnosing diseases more swiftly than most medical experts. There is a direct link between the insurer and the policyholder when the distance between an insurance business and the consumer is reduced to zero with the use of technology, especially digital health insurance. In comparison with traditional insurance, AI and machine learning have altered the way insurers create health insurance policies and helped consumers receive services faster. Insurance businesses use ML to provide clients with accurate, quick, and efficient health insurance coverage. This research trained and evaluated an artificial intelligence network-based regressionbased model to predict health insurance premiums. The authors predicted the health insurance cost incurred by individuals on the basis of their features. On the basis of various parameters, such as age, gender, body mass index, number of children, smoking habits, and geolocation, an artificial neural network model was trained and evaluated. The experimental results displayed an accuracy of 92.72%, and the authors analyzed the model's performance using key performance metrics.

Keywords: artificial intelligence; neural networks; machine learning; health insurance; prediction.

## 1. INTRODUCTION

We live in a world that is filled with dangers and uncertainties. People, homes, businesses, buildings, and property are all vulnerable to various types of risk, and these risks might differ. These threats include the risk of death, illness, and the loss of property or possessions. People's lives revolve around their health and happiness. However, because risks cannot always be avoided, the financial sector has devised a number of products to protect individuals and organisations from them by utilizing financial resources to compensate them. As a result, insurance is a policy that reduces or eliminates the expenses of various risks. A

*Corresponding Author

policy that protects medical bills is known as health insurance. An individual who has purchased a health insurance policy receives coverage after paying a certain premium. The cost of health insurance is determined by a variety of factors. The cost of a health insurance policy premium varies from person to person since various factors influence the cost of a health insurance plan. Consider age: a young individual is far less likely than an older person to suffer serious health issues. As a result, treating an elderly person is more expensive than treating a young one. As a result, an older individual must pay a higher premium than a younger person. Because [1] numerous factors influence the insurance premium of a health insurance policy, the premium amount varies from person to person. In healthcare, artificial intelligence is capable of completing many medical-related activities at a much quicker rate in order to forecast or diagnose illnesses/injuries effectively and deliver the best medical therapy to the patient. AI may gather data, process it, and offer the appropriate result to the user. This reduces the time it takes to detect diseases and mistakes, allowing the diagnosis–treatment–recovery cycle to be dramatically shortened. For example, if you choose an online consultation with a doctor, chatbots are used by healthcare professionals or organisations to obtain basic information prior to an appointment with the doctor. This assists the doctor in comprehending the problem before beginning the consultation procedure. As a result, both the doctor and the patient save time.

AI and ML play various roles in the health insurance market, some of which are listed below:

• The use of chatbots has become an increasingly important aspect of any firm; even healthcare organisations are embracing the technology. Because almost everyone has access to the Internet and a smartphone, interacting with physicians, hospitals, and insurance companies is much easier using chat applications. They are available 24 h a day, seven days a week, making them more effective than human interaction. They employ emotional analysis and natural language processing to better comprehend consumers' requests and respond to a variety of queries about insurance claims and product choices.

• Faster Claim Settlements: The time it takes for health insurance claims to be settled is one of the main difficulties for both policyholders and insurers. This might be due to lengthy manual processes or bogus claims. It takes time and effort to manually identify valid claims. However, AI has the potential to significantly lower claim processing times in the future. AI can detect fraudulent claims and learn from previous data to improve efficiency significantly.

• Personalised Health Insurance Policies: On the basis of an individual's past data and current health circumstances, insurers can identify and develop a health insurance plan for them. This assists the insurer in providing a proper health insurance plan rather than a health insurance package that clients may or may not utilise efficiently. Customers will also be urged to select a plan that meets their requirements rather than paying for services they may not use.

• Cost-effectiveness: Insurers are utilising AI to recommend good habits and behaviours to clients, such as exercise and diet, lowering the cost of avoidable healthcare expenditures caused by bad habits.

]• Fraud Detection: Researchers are working on building machines that can evaluate health insurance claims and anticipate fraud. This also aids insurers in resolving legitimate claims more quickly.

• Faster Underwriting: The health insurance underwriting procedure is lengthy and time-consuming. Fitness trackers, for example, can now collect and analyse vast amounts of data and share it with insurance companies thanks to technological breakthroughs, such as smart wearable technologies. Insurers can find innovative methods to underwrite consumers differently by employing these data. By adopting AI-based predictive analysis, health insurance firms may save time and money. Even as the healthcare business quickly digitises, enormous amounts of data will inevitably be created and gathered. This will simply increase the workload for healthcare providers since more raw data means more effort. For healthcare professionals and patients, AI can interpret these data and deliver insights based on them. It is a more efficient way to diagnose ailments. Some of the advantages of AI and ML in healthcare are:

• Clinical Observation-Based Decisions: AI and machine learning can process vast volumes of data in real time and give critical information that can aid in patient diagnosis and treatment recommendations. This translates to improved healthcare services at a reduced cost by evaluating patient data and delivering findings in a couple of minutes. Diabetes or blood sugar devices, for example, may analyse data rather than merely reading raw data and alert you to patterns depending on the information presented, allowing you to take immediate or corrective action.

## 2. RELATED WORK

In the life insurance sector, risk assessment is critical for classifying applicants. Companies utilise screening methodology to produce application decisions and determine the pricing of insurance products. The vetting process may be computerised to speed up applications or programs thanks to the expansion of data and advances in business intelligence. The goal of the study in [7] was to find ways to use predictive analytics to improve risk assessment for life insurance companies. The research was conducted using a real-world dataset with over a hundred characteristics (anonymised). Dimensionality reduction was performed to choose salient features that could increase the models' prediction potential. Actuaries utilise a variety of numerical procedures to forecast yearly medical claims expenditure in an insurance business. This sum must be accounted for in the annual financial budgets. Inaccurate estimation usually has a detrimental impact on a company's overall success. Goundar et al. [8] explained how to build an artificial neural network (ANN) that can predict yearly medical claims. The aim was to lower the mean absolute percentage error by changing factors of the configuration, such as the epoch, learning rate, and neurons, in various layers once the neural network models were constructed. Feed forward and recurrent neural networks were utilised to forecast the yearly claim amounts. Joseph Ejiyi et al. [9] investigated an insurance dataset from the Zindi Africa competition, which was stated to be from Olusola Insurance Company in Lagos, Nigeria, to demonstrate the efficacy of each of the ML algorithms we employed here. The results showed that, according to a dataset obtained from Zindi, insurance authorities, shareholders, administration, finance professionals, banks, accountants, insurers, and customers all expressed worry about insurance company insolvency. This worry stemmed from a perceived requirement to shield the general public from the repercussions of insurer insolvencies while also lowering management and auditing duties. In this work [10], we offer a strategy for preventing insurance company insolvency. In the past, insolvency prediction approaches, such as multiple regression, logit analysis, recursive partitioning algorithm, and others were applied. Fauzan and Murfi [11] used XGBoost to solve the issue of claim prediction and evaluate its accuracy. We also compared XGBoost's performance against that of other ensemble learning methods, such as AdaBoost, Stochastic GB, Random Forest, and Neural Network, as well as online learning methods. In terms of normalised Gini, our simulations suggest that XGBoost outperforms other techniques. People are increasingly investing in such insurance, allowing con artists to defraud them. Insurance fraud is a crime that can be committed by either the customer or the insurance contract's vendor. Unrealistic claims and post-dated policies, among other things, are examples of client-side insurance

fraud. Insurance fraud occurs on the vendor side in the implementation of regulations from non-existent firms and failure to submit premiums, among other things. In this study [12], we compare and contrast several categorisation methods.

Kumar Sharma and Sharma [13] aimed to develop mathematical models for predicting future premiums and validating the findings using regression models. To anticipate policyholders' choice to lapse life insurance contracts, we employed the random forest approach. Even when factoring in feature interactions, the technique beats the logistic model. Azzone et al. [14] studied how the model works; we employed global and local classification tools. The findings suggest that existing models, such as the logistic regression model, are unable to account for the variety of financial decisions. Understanding [15] the elements that influence a user's health insurance premium is critical for insurance firms to generate proper charges. Premium should always be a user's first concern when making suitable selections. The majority of characteristics that contribute to the cost of health care premiums are BMI, smoking status, age, and kids, according to the output, which revealed that these four parameters have a strong correlating effect on health insurance rates. Premiums are determined by health insurance companies' private statistical procedures and complicated models, which are kept concealed from the public. The goal of this study [16] is to see if machine learning algorithms can be used to anticipate the pricing of yearly health insurance premiums on the basis of contract parameters and business characteristics. The goal of this article [17] is to use a strong machine learning model to estimate the future medical costs of patients on the basis of specific parameters. Using the simulation results, the elements that influence individuals' medical expenditures were determined. The Japanese government has mandated that insurers develop a population health management strategy. To assess the strategy [18], a cost estimate is required. A standard linear model is not suited for the prediction since one insured patient might have several conditions. Using a quantitative machine learning technique, we created a medical cost forecasting model. The historical uniformity of health care expenses in a major state Medicaid programme was investigated in this research. The expenses were forecasted using predictive machine learning algorithms, particularly for high-cost, high-need (HCHN) patients. The findings of Yang et al. [19] indicated a high temporal link and showed potential for utilising machine learning to forecast future health care spending. HCHN patients had a stronger temporal association, and their expenditures may be anticipated more accurately. Including additional historical eras improved forecasting accuracy. Some individuals who are

economically disadvantaged will be unable to cover treatment-related fees. According to our behaviour and genetics, the necessity for health insurance varies as we grow older. Health insurance is becoming increasingly important as people's lifestyles and ailments change. Because a medical problem can strike anybody at any moment and have such a significant psychological and economic impact on the individual, it is difficult to predict when one will occur. With this background in mind, this research [20] aimed to forecast the cost of health insurance using the following contributions: age, gender, region, smoking, BMI, and children.

## 3. RESEARCH METHODOLOGY

In this paper, the authors used the Python programming language for the implementation and trained the machine learning-based model for the prediction of health insurance premiums. Initially, the dataset and the necessary python libraries and packages were imported. The dataset consisted of over 1300 entries and seven columns, namely charges, smoking, region, children, BMI, sex, and age. This dataset was used to predict the health insurance premium. Thereafter, an exploratory data analysis was performed. In this step, the dataset was checked for null values. Since there were no null values in the dataset, the statistical summary of the dataset was analysed. The statistical summary included the count, mean, standard deviation, and various other statistics related to the columns available in the dataset—age, BMI, number of children, and health insurance charges. The dataset link is given at the end of the paper in the Data Availability Statement.


Figure 1. Machine learning-based regression framework.

Regplot is a programme that plots data and fits a linear regression model. To evaluate the regression model, there are several mutually incompatible alternatives. In Figure 6, there is a straight line that passes through the data, and it seems that body mass index (BMI) tends to increase a little bit. It is possible that the charges also tend to increase slightly.

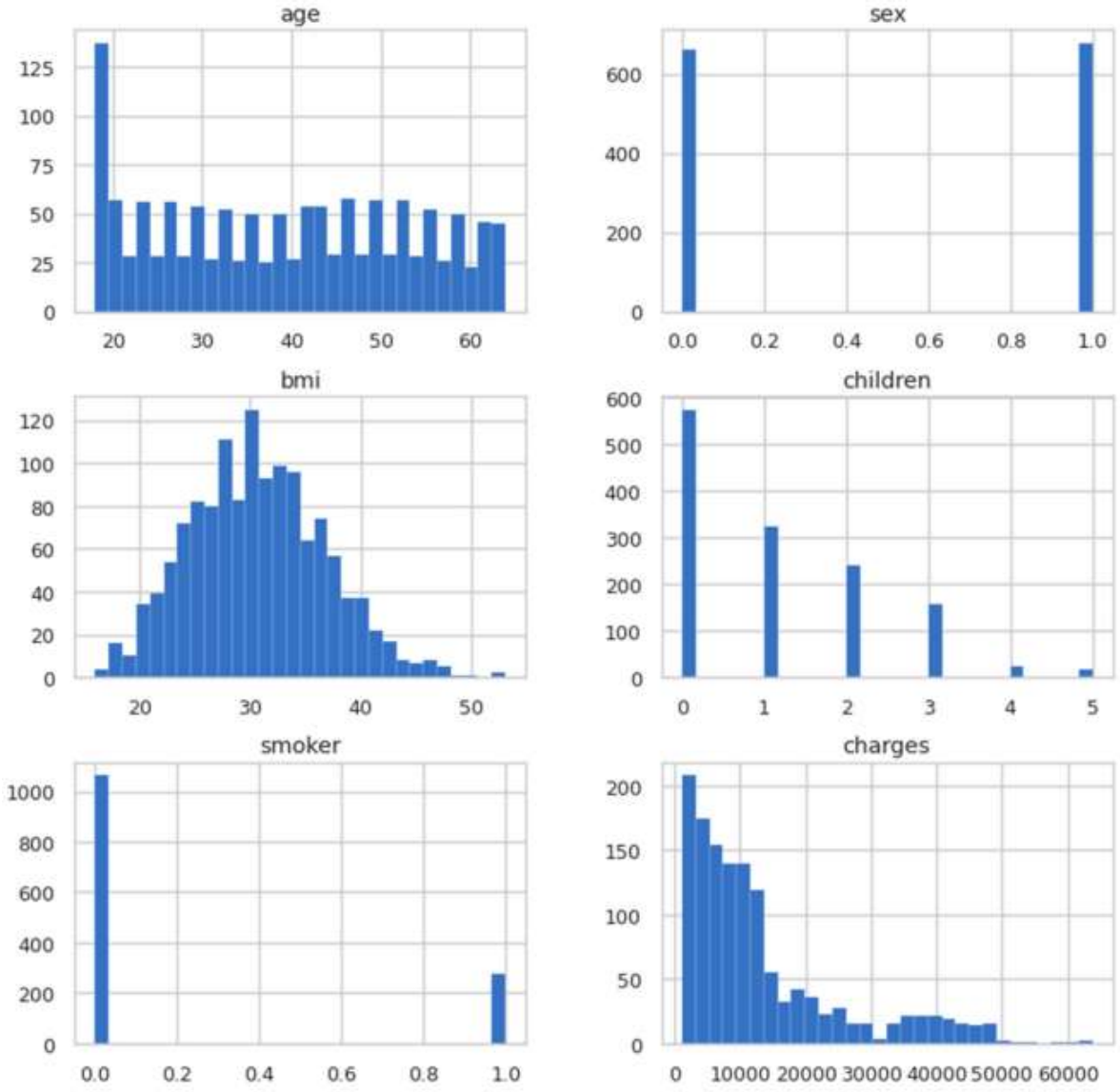


Figure 2. Histogram plots for columns.

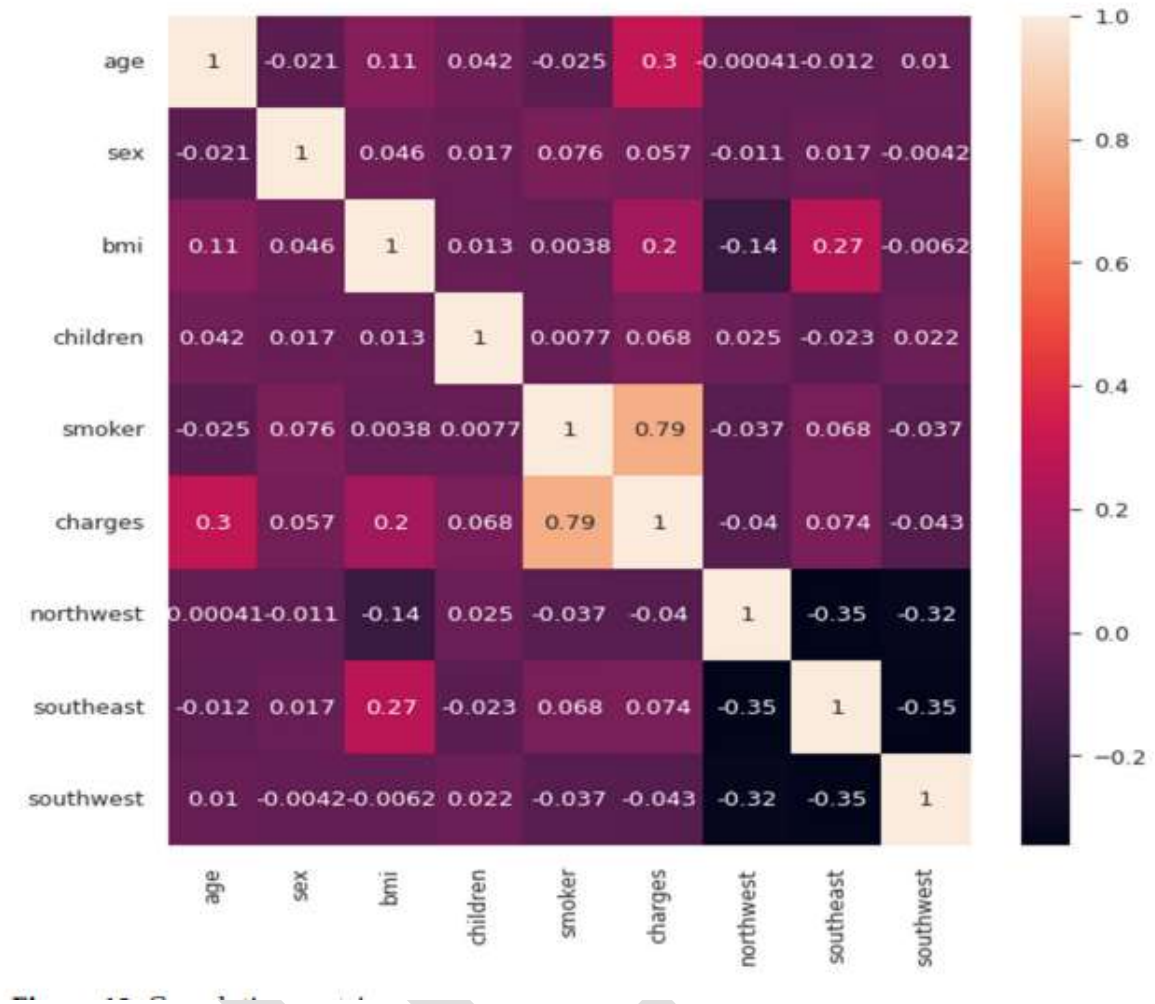


Figure 3. Correlation matrix.

## 4. CONCLUSIONS

In the field of health insurance, machine learning is well-suited to tasks that are often performed by people at a slower speed. AI and machine learning are capable of analysing and evaluating large volumes of data in order to streamline and simplify health insurance operations. The impact of machine learning on health insurance will save time and money for both policyholders and insurers. AI will handle repetitive activities, allowing insurance experts to focus on processes that will improve the policyholder's experience. Patients, hospitals, physicians, and insurance providers will benefit from ML's ability to accomplish jobs that are currently performed by people but are much faster and less expensive when performed by ML. When it comes to exploiting historical data, machine learning is one component of cognitive computing that may address various challenges in a broad array of

applications and systems. Forecasting health insurance premiums is still a topic that has to be researched and addressed in the healthcare business. In this study, the authors trained an ANN-based regression model to predict health insurance premiums. The model was then evaluated using key performance metrics, i.e., RMSE, MSE, MAE, r2, and adjusted r2. The accuracy of our model was 92.72%. Moreover, the correlation matrix was also plotted to see the relationship between various factors with the charges.

References

1. Health Insurance Premium Prediction with Machine Learning. Available online: https://thecleverprogrammer.com/2021/10/26 /health-insurance-premium-prediction-with-machine-learning/ (accessed on 9 May 2022).

2. ul Hassan, C.A.; Iqbal, J.; Hussain, S.; AlSalman, H.; Mosleh, M.A.A.; Sajid Ullah, S. A Computational Intelligence Approach for Predicting Medical Insurance Cost. Math. Probl. Eng. 2021, 2021, 1162553. [CrossRef]

3. Cevolini, A.; Esposito, E. From Pool to Profile: Social Consequences of Algorithmic Prediction in Insurance. Big Data Soc. 2020, 7. [CrossRef]

4. van den Broek-Altenburg, E.M.; Atherly, A.J. Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season. Appl. Sci. 2019, 9, 2035. [CrossRef]

5. Hanafy, M.; Mahmoud, O.M.A. Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models. Int. J. Innov. Technol. Explor. Eng. 2021, 10, 137–143. [CrossRef]

6. Bhardwaj, N.; Anand, R. Health Insurance Amount Prediction. Int. J. Eng. Res. 2020, 9, 1008–1011. [CrossRef]

7. Boodhun, N.; Jayabalan, M. Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. Complex Intell. Syst. 2018, 4, 145–154. [CrossRef] 8. Goundar, S.; Prakash, S.; Sadal, P.; Bhardwaj, A. Health Insurance Claim Prediction Using Artificial Neural Networks. Int. J. Syst. Dyn. Appl. 2020, 9, 40–57. [CrossRef]

[9]. Dr. B Sankara Babu, Srikanth Bethu, K. Saikumar, G. Jagga Rao, "Multispectral Satellite Image Compression Using Random Forest Optimization Techniques" Journal of Xidian University, in Volume 14, Issue 5-2020.

[10]. G. Jagga Rao, Y. Chalapathi Rao, "Human Body Parts Extraction in Images Using JAG-Human Body Detection (JAG-HBD) Algorithm Through MATLAB" Alochana Chakra Journal, Volume IX, Issue V, May/2020.

[11]. Dr. k. Raju, A. Sampath Dakshina Murthy, Dr. B. Chinna Rao,G. Jagga Rao "A Robust and Accurate Video Watermarking System Based On SVD Hybridation For Performance Assessment" International Journal of Engineering Trends and Technology (IJETT) – Volume 68 Issue 7 - July 2020.

[12]. G. Jagga Rao, Y. Chalapathi Rao, Dr. Anupama Desh Pande "A Study of Future Wireless Communication: 6G Technology Era " volume 14, issue 11,2020.

[13]. G. Jagga Rao, Y. Chalapathi Rao, Dr. Anupama Desh Pande "Deep Learning and AI-Based millimeter Wave Beamforming Selection for 6G With Sub-6 GHz Channel Information" Volume 21 : Issue 11 – 2020.

14. Azzone, M.; Barucci, E.; Giuffra Moncayo, G.; Marazzina, D. A Machine Learning Model for Lapse Prediction in Life Insurance Contracts. Expert Syst. Appl. 2022, 191, 116261. [CrossRef]

15. Sun, J.J. Identification and Prediction of Factors Impact America Health Insurance Premium. Master's Thesis, National College of Ireland, Dublin, Ireland, 2020. Available online: http://norma.ncirl.ie/4373/ (accessed on 9 May 2022). 16. Lui, E. Employer Health Insurance Premium Prediction. Available online: http://cs229.stanford.edu/proj2012/LuiEmployerHealthInsurancePremiumPrediction.pdf (accessed on 17 May 2022).

17. Prediction of Health Expense—Predict Health Expense Data. Available online: https://www.analyticsvidhya.com/blog/2021/0 5/prediction-of-health-expense/ (accessed on 9 May 2022).

18. Takeshima, T.; Keino, S.; Aoki, R.; Matsui, T.; Iwasaki, K. Development of Medical Cost Prediction Model Based on Statistical Machine Learning Using Health Insurance Claims Data. Value Health 2018, 21, S97. [CrossRef]

19. Yang, C.; Delcher, C.; Shenkman, E.; Ranka, S. Machine Learning Approaches for Predicting High Cost High Need Patient Expenditures in Health Care. Biomed. Eng. Online 2018, 17, 131. [CrossRef] [PubMed]

20. Shyamala Devi, M.; Swathi, P.; Purushotham Reddy, M.; Deepak Varma, V.; Praveen Kumar Reddy, A.; Vivekanandan, S.; Moorthy, P. Linear and Ensembling Regression Based Health Cost Insurance Prediction Using Machine Learning. Smart Innov. Syst. Technol. 2021, 224, 495–503.