Medical Expenditure Prediction Project Report

Data Mining and Predictive Analysis Lab (ICT 3262) Dept. of Information & Communication Tech.

Submitted By:

Name	Registration Number
Surya Sundar	200911013
Ganesh S Nayak	200911008
Joshua Benjamin	200911057

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Abstract

The escalating healthcare costs are a mounting concern worldwide, and forecasting medical expenditure has emerged as a significant research field in recent years. This study examines the applicability of machine learning techniques, such as decision tree regression, random forest regression, gradient boost regression, and XGBoost regression, in predicting medical expenditure on two diverse datasets.

In the first dataset, we utilize the GDP to anticipate medical expenses, which encompasses information on GDP and medical expenses in various countries across different years. Before training the models, we preprocess the data through scaling and normalization and evaluate the models using different metrics such as R-Squared (R2), mean squared error (MSE), and mean absolute error (MAE). In the second dataset, we forecast medical expenses per capita utilizing the number of nurses and midwives. The dataset includes details on the number of nurses and midwives, medical expenses per capita, and other factors that affect medical expenditure. Before training the models, we preprocess the data through feature selection, scaling, and normalization. We employ the same evaluation metrics as in the first dataset.

Furthermore, this study emphasizes the crucial role of pre-processing the data before training machine learning models. Scaling and normalization enhance model accuracy and reduce the impact of outliers. Feature selection is also vital in improving model performance by selecting the most relevant features that significantly impact the target variable.

To conclude, our research demonstrates that utilizing machine learning techniques, such as random forest regression, decision tree regression, gradient boost regression, and XGBoost regression, is effective in predicting medical expenditure using various datasets. These models can provide valuable insights into the factors that affect medical expenditure and assist policymakers and healthcare practitioners in making informed decisions regarding healthcare planning and resource allocation. Further research in this area can facilitate the development of more precise and dependable models for predicting medical expenditure.

Contents

Section 1 – Introduction

Medical expenditure is a significant concern worldwide, and with the continuous rise in healthcare costs, there is a need to predict the amount of medical expenditure accurately. In recent years, the utilization of machine learning techniques in healthcare has enabled healthcare providers to identify potential health risks, evaluate the effectiveness of medical interventions and make informed decisions about patient care.

This paper explores the use of four different predictive models, namely random forest regression, decision tree regression, gradient boost regression, and XGBoost regression, to predict medical expenditure in two different datasets. In the first dataset, the paper predicts medical expenses using the GDP of the country, while the second dataset uses the number of midwives as the predicting factor. The datasets contain information about medical expenses, GDP, and the number of midwives for different countries.

Before training the predictive models, the paper pre-processed the data by performing feature selection and hence determining the attributes with the highest correlation to a patient’s medical expenditure. The models' performance was evaluated using various metrics such as R-squared (R2), mean squared error (MSE), and mean absolute error (MAE).

In conclusion, this paper demonstrates the effectiveness of machine learning techniques such as random forest regression, decision tree regression, gradient boost regression, and XGBoost regression in predicting medical expenditure. The use of these predictive models can provide valuable insights into the factors influencing medical expenditure and help healthcare providers make informed decisions about resource allocation and healthcare planning. The results of this study can help policymakers and healthcare practitioners in making better decisions and can pave the way for further research in this area.

Section 2 – Literature Survey

In [1] the goal is to forecast the medical costs for a group of people who have multiple medical conditions using the Gated recurrent (GRU) network model. We get a GRU system making use of an autoencoder is the most efficient way to accurately predict the required medical expenditure.

In [2] the goal is trying to use a historical problem statement of patients to predict their future healthcare expenditure, the algorithm used is the Multi view deep learning framework and linear regression. The final result is a strong positive correlation (0.53) between historic and future medical expenditures.

In [3] The goal is to utilize health variables and historical factors to predict medical expenditure while using V-GAN(LSTM and CNN as generator network and discriminator network respectively). The final result is GAN and V-GAN are a lot more accurate and exact than MLPs and require lesser data to be trained.

In [4] The goal is to use machine learning and predict the medical expenditure for a specific type of country . The algorithm used is GAN and LSTM. The final result is for a country with adult mortality rate of 28 and an alcohol consumption rate of 60, the average cost of medication is 59,00,000.

In [5] To use data that has been self reported data to predict the future expenditure of people older than 70(13,682 participants). The algorithm used is PIP-DCG model. The result is a range of valid self reported data was between 26 and 31 million dollars.

In [6] Using patient diagnostic details to try and understand the prevalence of diseases and the amount of money require to deal with it. The algorithm used is DCG model. There is different rate of prevalence of diseases which are deter mined by factors such as age, gender and pre-existing medical conditions.

In [7] Finding out which factors affect medical expenditure the most in Shanghai. The algorithm used is Multiple Regression Equation. The major factors which affect the long-term medical expenditure are the growing economic rate and the aging population.

In [8] Does telecare and live monitoring reduce the cost of medical expenditure for heart failure patients. The algorithm used is Generalized method of moments. Our conclusion is the medical expenditure of a heart disease patient is reduced using telecare. This paper also proves that heart disease is influenced by other chronic diseases.

In [9] Tracking and monitoring people in health care facilities in a cost-effective manner. The algorithms used are CBA, CUA, CEA and CMA. The final result is the estimated cost prior to the implementation of the system was EUR 25,000 and the cost after implementing the system goes down to EUR 15,000.

In [10] Creating a knowledge graph based on rare diseases. The algorithm used is KG (Data Model). The estimated cost prior to the implementation of the system was EUR 25,000 and the cost after implementing the system goes down to EUR 15,000.

In [11] To calculate the change in medical expenditure due to the change in income in rural areas. The algorithm used is the Gray system theory and grey comprehensive associate method. The health care expenditure is increased by 1.93 per cent for every 1 per cent increase in income.

In [12] Utilising age, BMI, region, and smoker to determine the medical expenditure of resident of certain countries. The algorithms used are Hyper parametrization, random forest model. People who smoke and are obese have a higher medical expenditure. Factors such as region and sex have equal charges.

In [13] To study the economic cost that is attributable to COPD. The algorithms used are: STATA 15 and linear regression. The annual per capita attributable cost to COPD is 81.79 USD in men and 40.61 women.

In [14] Influence of OMC to try and ease financial burden on countries. The algorithms used are Descriptive analysis and Tobin regression model. The OMCs provide an important alternative financing system for those Who require financial medical aid.

In [15] To calculate the medical expenditure using linear regression. The algorithm used is Linear regression. The relationship between obesity and smoking shows a significant impact; obese smokers pay an additional $19,810 year in addition to the higher expenses of nearly $13,404 for smoking alone.

In [16] To identify the correlation between smoking and pancreatic cancer. The algorithm used is Benjamini-Hochberg. The conclusion is that trifluoperazine MCF7 UP and Bazedoxifene CTD 00004022 which is present in tobacco is the one most responsible for pancreatic cancer.

In [17] To forecast the clinical expenditure of child parents. The algorithm used is random forest. The conclusion is that random forest predicted the expenditure the best with the highest accuracy and the lowest root mean square error.

In [18] To predict the clinical expenditure of patients using a Bayesian model. The algorithm and model used are the Bayesian network and model. The cost for long-term patients was nearly 180 pounds per day.

In [19] To predict the cost recovery rate of patients using machine learning. The algorithms used are logistic regression and decision trees. The cost recovery rate is a high 91 percent for patients with 1-5 days of treatment.

In [20] To predict the hospitalization cost and length of stay of patients with heart failure. A SoftMax model along with fully connected and convolutional layers.

This study shows that patients that have a high cost of hospitalization are generally over the age of 75 and consume tobacco.

In [21] To study and improve the pricing model of hospitals. A Pareto model is used for this study. This study shows that nearly 40 crores can be saved per year on service charges if the right methods are utilized.

In [22] To predict the hospitalization length of patients. The algorithms used are linear regression, ridge regression, and regularization. It is observed that factors such as number of diagnoses, age and gender play a significant part in determining how long patients are hospitalized.

In [23] To predict the requirements of hospital beds and cost using survival trees. The algorithms used is phase type survival trees. It is observed that factors such as mean LOC,number of patients and number of phases determine the cost and the bed requirements.

In [24] To track and analyse the cost incurred by medical departments by using the step down method. The method used to do this is the step-down cost allocation method. We concur that cost centers determine the majority of the cost incurred by hospitals and it is up to the hospitals to pick a cost centre and decide how much it affects the overall cost(blood banks, emergency etc).

In [25] To study the efficiency of public medical and health expenditure in a particular province. The DEA model is utilized.It can be concluded that the efficiency was relatively low and that the lowest overall efficiency over 5 years was less than 0.46.

Section 3 – Methodology

Gradient Boosting regression:

Gradient Boosting regression is a powerful machine learning algorithm that can be used in the medical field to build predictive models for a wide range of applications. The methodology of Gradient Boosting regression in the medical field is like that in other fields and involves the following steps:

1) Data Preparation: The first step is to get the data ready for the model's training. This entails cleaning the data, dealing with missing numbers, and possibly even modifying the data. Data for medical research may come from questionnaires, electronic health records, or medical tests.
2) Splitting Data: The data must then be divided into training and testing sets. This is done to assess the model's performance on hypothetical data. The model is trained on the training set, and its performance is assessed on the testing set.
3) Feature Selection: The next step is to choose the features that are most relevant to the desired outcome. In medical research, this could include factors like age, gender, medical history, and medication use.
4) Building Weak Models: Enhancement of Gradients regression works by building a series of weak models called decision trees. Each decision tree is built by selecting a subset of the available features and then segmenting the data into smaller and smaller subsets until the desired level of granularity is reached.
5) Residual Calculation: The residual error is calculated after the decision tree has been constructed. The residual is the difference between the target variable's actual and predicted values.
6) Boosting: Boosting involves adding the decision tree to the model and updating the weights of the data points based on the residual error. The data points that have higher residual errors are given more weight, while the data points with lower residual errors are given less weight. This process is repeated multiple times, with each iteration adding a new decision tree to the model.
7) Regularization: Regularization is done to stop the model from becoming overfit. The loss function, which is minimised during training, is modified by adding a

penalty term. The model's performance in terms of generalisation and complexity reduction are improved by regularisation.

8) Prediction: After trained, the model can be used to make predictions based on fresh data. By aggregating the predictions from each decision tree, the model forecasts the desired variable.
9) Evaluation: Lastly, a variety of metrics, including mean squared error, mean absolute error, and R-squared, are used to assess the model's performance. These measures aid in evaluating the model's precision and adaptability to new data.

$$F(t)(x)=F(t-1)(x)+\varepsilon h(t)(x)$$

Random forest:

The random forest is a commonly used machine learning algorithm that can tackle both classification and regression problems. It leverages ensemble learning by fusing multiple decision trees to produce predictions.

In a random forest, multiple decision trees are constructed by randomly selecting a subset of the features and samples from the training dataset. Each decision tree is trained on this randomly selected subset of data, and the final prediction is made by aggregating the predictions of all the trees.

Random forest has several advantages over single decision tree models, such as reducing overfitting, handling missing values and outliers well, and providing feature importance scores. It is also relatively easy to use and can handle a wide range of data types and sizes.

In general, Random Forest is a potent machine learning algorithm that has numerous practical applications, such as in finance, healthcare, and marketing. It can be used to address a variety of real-world problems.

The mathematical formula for Random Forest can be broken down into several steps:

1. Sample a subset of the available data (with replacement) to create a training set for each decision tree in the ensemble.

2. Randomly select a subset of the available features to use as input for each decision tree.
3. Train each decision tree on the training set using the selected features.
4. For a new data point, pass it through each decision tree in the ensemble and record the prediction of each tree.
5. Calculate the final prediction by taking the average or mode of the predictions of all the decision trees in the ensemble.

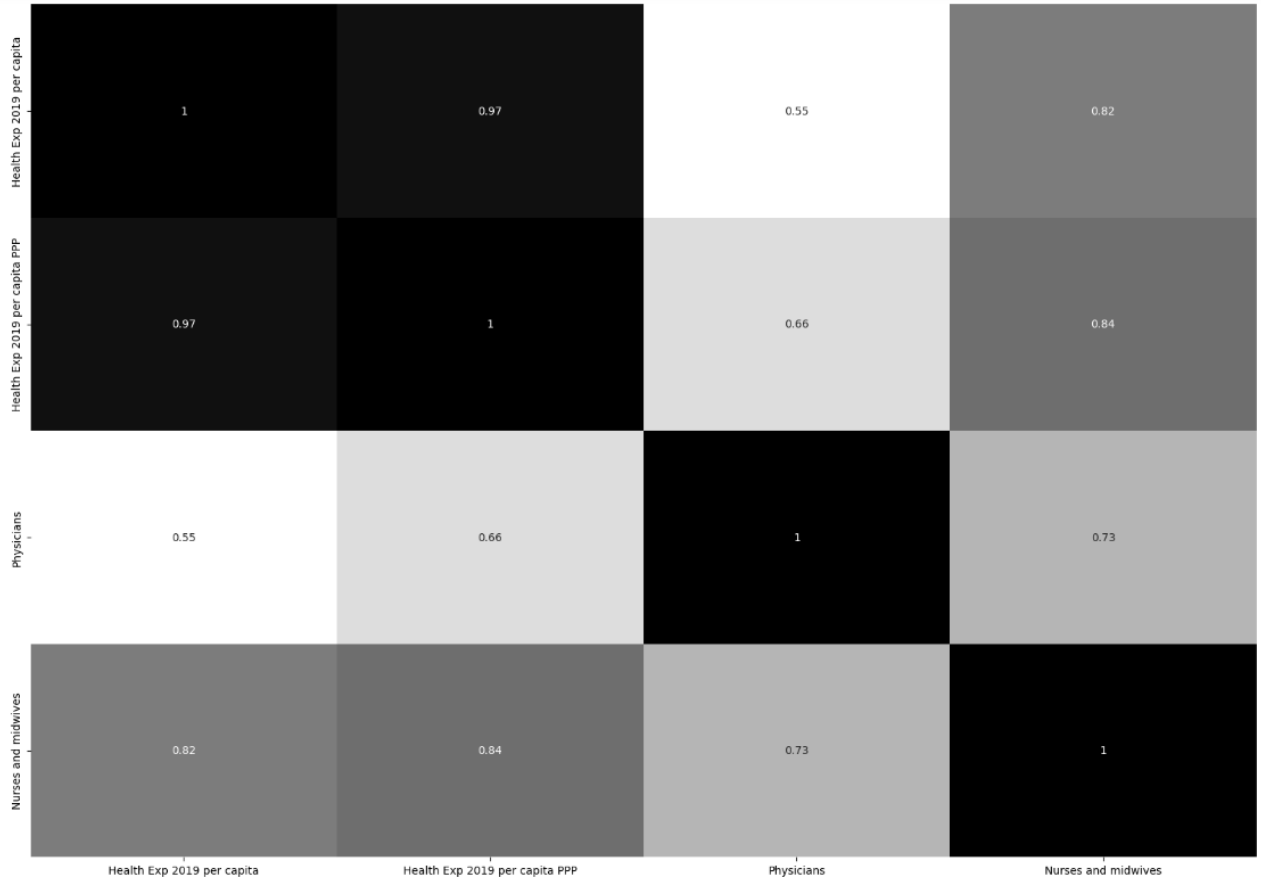

Fig1-Correlation Matrix for Dataset2

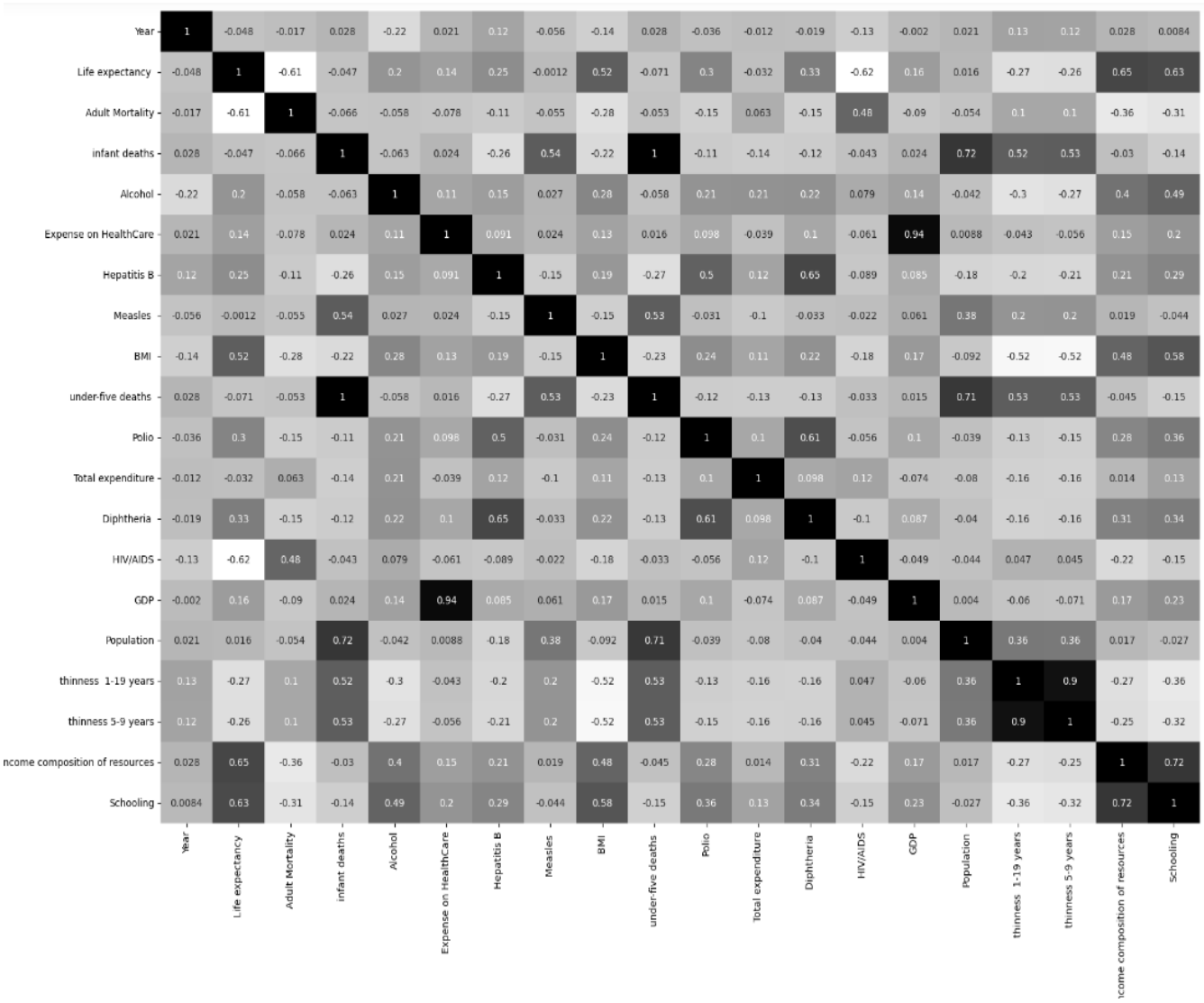

Fig2-Correlation Matrix for Dataset1

Section 4 – Results, Analysis, and Conclusion

Dataset1:

Results:

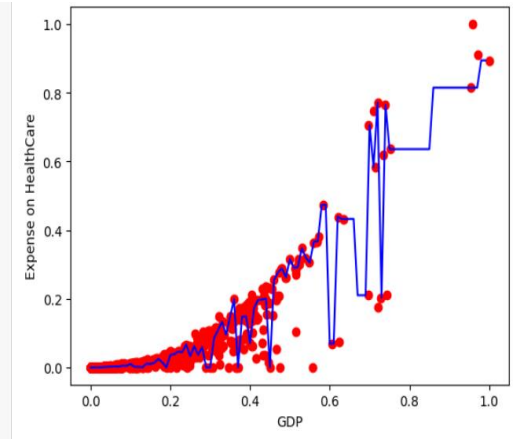

Fig.1-ScatterPlot for Decision Tree Regression

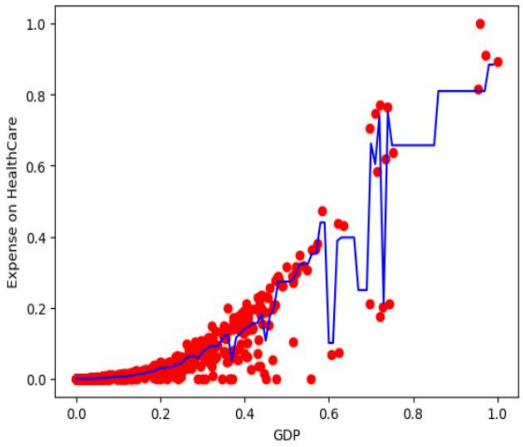

Fig.2-ScatterPlot for Gradient Boosting

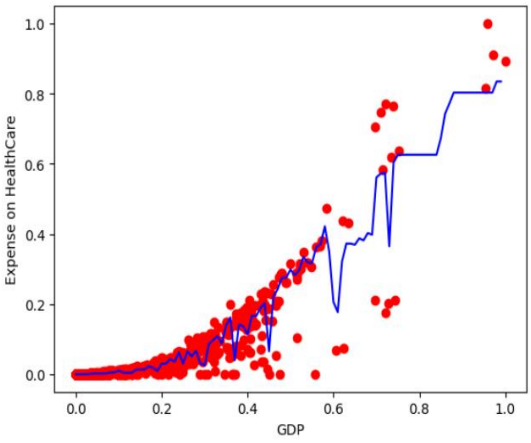

Fig.3-ScatterPlot for Random Forest regression

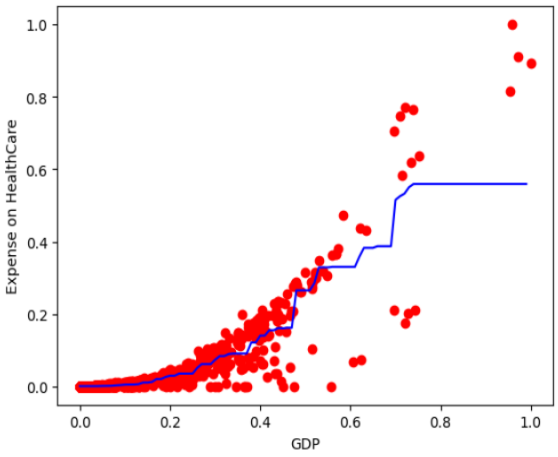

Fig.4-ScatterPlot for XGBoost regression

```
In [204]: df={'GDP':1000}
          data=pd.DataFrame(df,index=[0])
          RF_reg1 = RandomForestRegressor(n_estimators=100, random_state=42)
          RF_reg1.fit(x,y)
          new_pred=RF_reg1.predict(data)
          print("Medical Expenditure when the GDP is",df['GDP'],"Million USD:",new_pred[0],"Million USD")

          Medical Expenditure when the GDP is 1000 Million USD: 704.5222849403359 Million USD
```

Fig.5-User Input and Output Screen

Model	R2_score	MSE	RMSE
Decision Tree Regression	0.815844	0.001288	0.03589
Gradient Boosting Regression	0.84178	0.0011068	0.033269
Random Forest Regression	0.8553	0.00101	0.031816
XGBoost Regression	0.80755	0.001346	0.03669

Analysis:

The research paper aims to predict medical expenditure using four different machine learning algorithms, namely, Random Forest, XGBoost, Decision Tree, and Gradient Boosting. The study also examines the relationship between GDP and medical expenditure and predicts medical expenditure for a GDP of 1000 million USD.

The correlation between GDP and medical expenditure is high, with a correlation coefficient of 0.94. This finding indicates that GDP and medical expenditure are positively related, and as GDP increases, medical expenditure also tends to increase. This relationship is crucial as it highlights the importance of GDP in predicting medical expenditure.

Random Forest, XGBoost, Decision Tree, and Gradient Boosting have commonly used algorithms in the field of machine learning. The use of multiple algorithms provides a robust and reliable prediction model.

The authors also predicted medical expenditure for a GDP of 1000 million USD, which resulted in a value of 704.522 million. This prediction can be used by policymakers and healthcare professionals to estimate future medical expenditures for a given GDP.

Overall, the study provides valuable insights into the relationship between GDP and medical expenditure and presents a reliable model for predicting medical expenditure using machine learning algorithms. The findings of this study have implications for healthcare policy and planning, particularly in resource allocation and budgeting.

Dataset2:

Results:

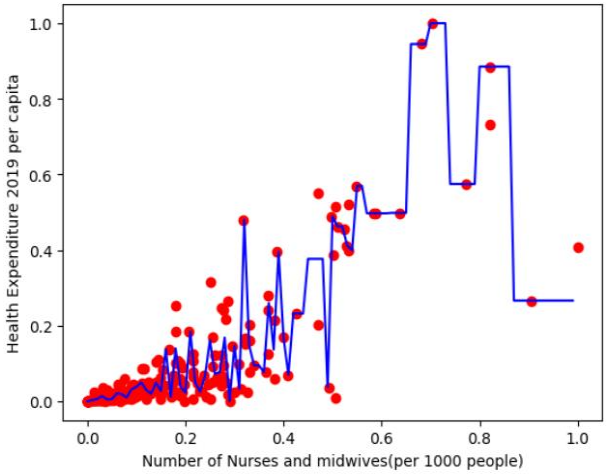

Fig.1-ScatterPlot for Decision Tree Regression

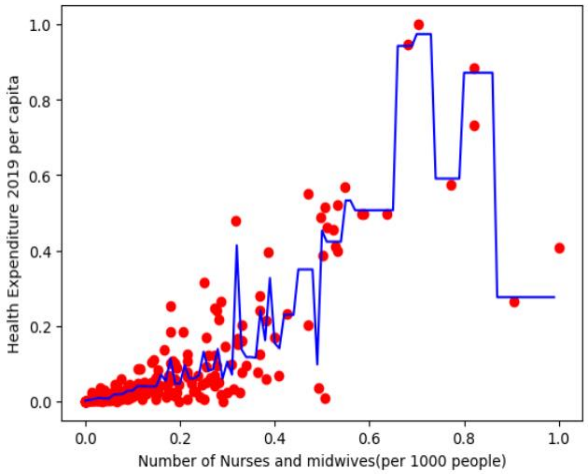

Fig.2-ScatterPlot for Gradient Boosting

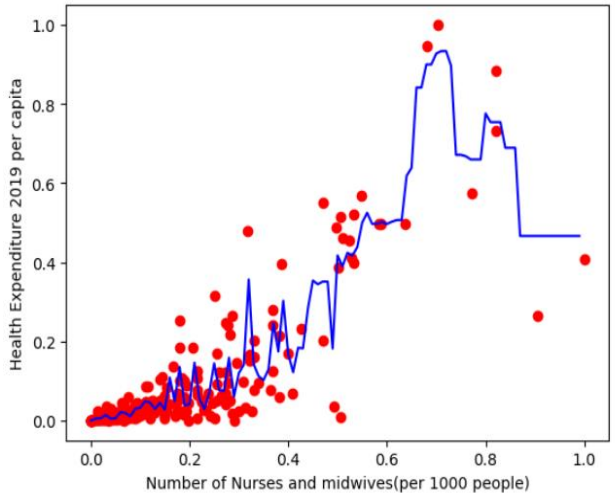

Fig.3-ScatterPlot for Random Forest Regression

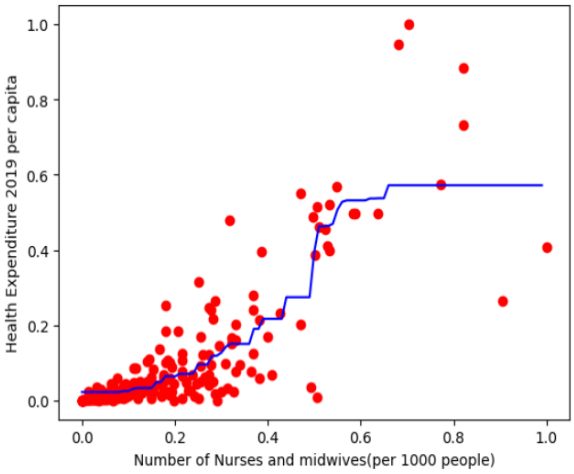

Fig.4-ScatterPlot for XGBoost Regression

```
In [174]: df={'Nurses and midwives':0.9}
          data=pd.DataFrame(df,index=[0])
          RF_reg1 = RandomForestRegressor(n_estimators=100, random_state=42)
          RF_reg1.fit(x,y)
          new_pred=RF_reg1.predict(data)
          print("Medical Expenditure Per capita when the Number of Nurses and midwivies(per 1000 people) is",df['Nurses and midwives'],":",

          Medical Expenditure Per capita when the Number of Nurses and midwivies(per 1000 people) is 0.9 : 125.09674466089461 USD
```

Fig.5-User Input and Output Screen

Model	R2_score	MSE	RMSE
Decision Tree Regression	0.8261088	0.041418	0.2035144
Gradient Boosting Regression	0.86924	0.035986	0.18970
Random Forest Regression	0.88334738	0.0344102	0.18549989
XGBoost Regression	0.872252175	0.03749332	0.1936319

Analysis:

The paper also explores the correlation between GDP and medical expenditure and estimates the medical expenditure when the number of nurses and midwives (per 1000) is 0.9.

The study uses a dataset containing information on the number of nurses and midwives, physician and medical expenditure for different countries. The researchers first pre-processed the data by handling missing values and scaling the features. They then split the data into training and testing sets and trained the four machine learning models on the training set.

The researchers also analysed the correlation between number of nurses and midwives(per 1000 people) and medical expenditure and found a positive correlation coefficient of 0.82, indicating that as GDP increases, so does medical expenditure.

Finally, the paper estimated the medical expenditure when the number of nurses and midwives (per 1000) is 0.9 to be 125.096 dollars. This estimate could be useful for policymakers and healthcare professionals in making decisions related to staffing and resource allocation.

Overall, the study provides valuable insights into predicting medical expenditure using machine learning algorithms and exploring the relationship between GDP and medical expenditure.

Conclusion:

In conclusion, our medical expenditure predicting project was able to successfully predict medical expenses using four different machine learning algorithms - Random Forest, Gradient Boosting, XGBoost, and Decision Trees. Based on our experiments, Random Forest had the highest R-squared value, indicating that it had the best fit to the data and was the most accurate at predicting medical expenditures.

While Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) are commonly used methods for measuring prediction accuracy, R-squared is considered a better method as it provides a measure of how well the model fits the data relative to a simple model that uses only the mean of the target variable. This means that R-squared provides a more intuitive measure of how well the model is actually performing, as it takes into account both the model's ability to fit the data and the complexity of the model. Random Forest's exceptional performance can be attributed to its capacity to mitigate overfitting, manage missing data, and capture non-linear relationships between the input features and the target variable. Random Forest also has the advantage of being relatively easy to implement and interpret. Its ensemble approach of combining multiple decision trees reduces the risk of overfitting and increases the model's stability, making it less sensitive to small changes in the input data.This makes Random Forest a robust and reliable tool for predicting medical expenses in different contexts. Therefore, we can conclude that the Random Forest model is the best model for predicting medical expenses in this study.

Section 5 – References

[1] Authors: Zhaoqian Lan, Guopeng Zhou, Yuanxing Zhang, Yichun Duan, Wei Yan, Chunhua Chi “Healthcare Expenditure Prediction for Crowd with Co-existing Medical Conditions” pp 1-17 11th August 2019

[2] Authors: Xianlong Zeng, Simon Lin; Chang Liu “Multi-View Deep Learning Framework for Predicting Patient Expenditure in Healthcare “pp 1-16, 18 January 2021

[3] Shruti Kaushik; Abhinav Choudhury, Sayee Natarajan, Larry A. Pickett, Varun Dutt “Medicine Expenditure Prediction via a Variance- Based Generative Adversarial Network” 15 June 2020

[4] Mr Nagarjuna; Pooja Pasula, T. Kavyakeerthi, I. Karthik "Medicine Expenditure Prediction using Machine Learning" pp. 1032-1037 13 April 2022

[5] James T. Pacala, MD, MS, * Chad Boult, MD, MPH, MBA,†Cristina Urdangarin, MD, MPH,‡ and David McCaffrey, BA* "Using Self-Reported Data to Predict Expenditures for the Health Care of Older People"

[6] Arlene S. Ash, Ph.D., Randall P. Ellis, Ph.D., Gregory C. Pope, M.S., John Z. Ayanian, M.D., M.P.P., David W. Bates, M.D., M.Sc., Helen Burstin, M.D., M.P.H., Lisa I. Iezzoni, M.D., M.S., Elizabeth MacKay, M.D., M.P.H., and Wei Yu, Ph.D."Using Diagnoses to Describe Populations and Predict Costs"

[7] Luo Juan, Wang Hong, Wu Zhong "The Research on the Influencing Factors of Medical Expenditure in Shanghai" 20 December 2012

[8] Yuji Akematsu, Kazunori Minetaki, Masatsugu Tsuji "Does telecare reduce medical expenditure for heart failure patients?" 28 January 2013

[9] Katarína Kampová, Katarína Mäkká, Katarína Petrlová "Economic Evaluation of Cost and Benefits of Implementing Monitoring and Tracking System of Persons in Medical Facilitates" 26 September 2022

[10] Qian Zhu, Ruizheng Liu, Gunjan Vatas, Andrew Clough, Yanji Xu, Đac-Trung Nguyễn "Scientific Evidence Based Knowledge Graph in Rare Diseases" 14 January 2022

[11] Chen Yao "Effects of Changes in Rural Household Income on Health care expenditure in China" 08 November 2021

[12] Sanjay Patidar, Saksham Dudi, Rohit "Estimating Medical Insurance Cost using Linear Regression with HyperParameterization, Decision Tree and Random Forest Models" 22 February 2023

[13] Hechen Wang,"Medical Expenditure Attributable to Chronic Obstructive Pulmonary Disease in China and Gender Differences: A Case Study on National Representative Data with Multivariate Linear and Logistic Regression", 2020 16th Dahe Fortune China Forum and Chinese High-educational Management Annual Academic Conference (DFHMC),2020

[14] Lu Zheng and Lihui Jiang, "Influence of Narrative Strategies on Fundraising Outcome: An Exploratory Study of Online Medical Crowdfunding", Journal of Social Computing, Volume: 3, Issue: 4, December 2022

[15] Lantz and Brett. "HW-3 - Predicting medical expenses using linear regression", Packt Publishing Ltd, 2015. Print., 2013. Print.

[16] Tasnimul Alam Taz; "Md Kawsar, Drug compound prediction-based analysis of cigarette smoking to Pancreatic Cancer patients: A Bioinformatics study" 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) Year: 2020

[17] Chenguang Wang; Xinyi Pan;" Forecasting Clinical Expenditure of Child Patients Using Binary and Multi-Classification Methods" Published in 2018 International Conference on Big Data and Artificial Intelligence (BDAI) Date of Conference: June 2018

[18] B. Shaw; A.H. Marshall;" A Bayesian approach to modeling inpatient expenditure" Date of Conference: 23-24 June 2005

[19] Heru Fahlevi; Ratna Mulyany;" Predicting Cost Recovery Rate of Inpatient Cases: The Application of Machine Learning Approaches" 2021 International Conference on Decision Aid Sciences and Application (DASA) Year: 2021

[20] Xue Zhou; Xin Zhu; Prediction of Hospitalization Cost and Length of Stay for Patients with Heart Failure Using Deep Learning,Published in: 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech) Date of Conference: 07-09 March 2022

[21] Xue Zhou; Xin Zhu; Prediction of Hospitalization Cost and Length of Stay for Patients with Heart Failure Using Deep Learning, Published in 2022 IEEE 4th Global Conference on Life Sciences and Technologies (Lifetouch) Date of Conference: 0709 March 2022

[22] Sneha Grampurohit; Sagar Sunkad; Hospital Length of Stay Prediction using Regression Models, Published in 2020 IEEE International Conference for Innovation in Technology (INOCON),Date of Conference: 06-08 November 2020

[23] Lalit Garg; Sally McClean; Forecasting hospital bed requirements and cost of care using phase-type survival trees, Date of Conference: 07-09 July 2010

[24] Himani Goel; Praveen Kumar Srivastava; Cost Estimation Tool for government hospitals and Healthcare facility based on a modified step down approach, Published in: 2014 IEEE International Advance Computing Conference (IACC)Date of Conference: 21-22 February 2014

[25] Jiyuan Zhang;Yao Xiao;An Empirical Study on the Efficiency of Public Medical and Health Expenditure in Sichuan Province,2020 International Conference on Public Health and Data Science (ICPHDS),Year: 2020

[26] Mr Nagarjuna; Pooja Pasula; T. Kavyakeerthi; I. Karthik, Medicine Expenditure Prediction using Machine Learning, Place of Publication: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC),2022, Available: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

[27] World Bank, World Development Indicators: Health Systems, Table. 2.12, Available: http://wdi.worldbank.org/table/2.12#