International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-15 Issue-4, September 2025

Application of Machine Learning Models for Patients Health Insurance Cost Prediction

Annwesha Banerjee Majumder, Sumit Das, Aniruddha Biswas, Trishita Ghosh, Raj Poddar, Suchetana Chakraborty

Abstract: ***The use of machine learning models to forecast health insurance costs based on personal characteristics is examined in this study. Age, sex, BMI, number of children, smoking status, and region were among the demographic variables included in the dataset. It was investigated how well several machine learning methods, such as Random Forest, Gradient Boosting, and Linear Regression, estimated insurance costs. After preprocessing the dataset by scaling numerical features and encoding categorical variables, k-fold cross-validation was employed to train and evaluate the regression models. The coefficient of determination (R2), mean absolute error (MAE), and root mean squared error (RMSE) were used to evaluate performance. According to experimental results, Gradient Boosting performed better than Random Forest and Linear Regression.***

Keywords: Gradient Boosting. Linear Regression, Mean Squared Error, Random Forest, Root Mean Squared Error.

Abbreviations:
RMSE: Root Mean Squared Error
MAE: Mean Absolute Error
GBMs: Gradient Boosting Machines
CNNs: Convolutional Neural Networks
RL: Reinforcement Learning
NLP: Natural Language Processing
EHRs: Electronic Health Records
ML: Machine Learning

I. INTRODUCTION

The ability to estimate and analyse medical expenditures has become increasingly critical in the modern healthcare industry due to the need for effective risk management and efficient resource allocation.

Manuscript Received on 05 August 2025 | Revised Manuscript Received on 06 September 2025 | Manuscript Accepted on 15 September 2025 | Manuscript published on 30 September 2025.
*Correspondence Author(s)
Dr. Annwesha Banerjee Majumder*, Assistant Professor, Department of Information Technology, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: annwesha.banerjee@jiscollege.ac.in
Dr. Sumit Das, Associate Professor, Department of Information Technology, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: sumit.das@jiscollege.ac.in
Aniruddha Biswas, Assistant Professor, Department of Information Technology, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: aniruddha.biswas@jiscollege.ac.in
Trisita Ghosh, Assistant Professor, Department of Information Technology, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: trishita.ghosh@gnit.ac.in
Raj Poddar, Department of Information Technology, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: rajpoddar8907@gmail.com
Suchetana Chakraborty, Department of Information Technology, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: chowdhury.mizan@jisuniversity.ac.in

By utilising patient data, including demographics, medical history, and prescribed medications, machine learning models have become practical tools for predicting medical costs. These models provide precise projections of future healthcare costs by utilising sophisticated algorithms, including Linear Regression, Random Forest, and Gradient Boosting. These predictive skills offer valuable insights, enabling insurance companies, governments, and healthcare providers to make informed decisions.

These models enable more efficient resource allocation and the creation of personalised insurance plans by identifying high-cost patients and facilitating targeted interventions. Through improved cost control and budget planning, they also significantly contribute to the overall efficiency of healthcare systems. Beyond the monetary consequences, applying machine learning to medical cost prediction enhances patient care by providing data-driven insights that facilitate individualised treatment plans. This study examines the approaches, uses, and effects of machine learning in medical cost forecasting, emphasising its potential to revolutionise the healthcare industry.

In this work a medical cost prediction model has been proposed applying different machine learning methods and their performances have been analyzed in details. The primary objective of this research is to develop a reliable medical cost prediction model by employing a range of machine learning approaches, with a focus on evaluating the effectiveness and performance of each strategy. The project aims to support stakeholders, including governments, healthcare providers, and insurance companies, by accurately predicting healthcare expenses using patient data, such as demographics, medical history, and specific medications. Using K-fold cross-validation, it also aims to examine and compare the effectiveness of Random Forest and Gradient Boosting regression models, while providing practical guidance for risk management, resource allocation, and the development of personalised insurance plans.

Random Forest Regressor and Gradient Boosting Regressor have been used with the K-Fold cross-validation technique. The dataset used in the model was collected from Kaggle. Label encoding has been applied to the dataset to convert numerical data into categorical data.

This study makes significant contributions in several ways. By comparing machine learning approaches, it first creates and evaluates a model for predicting medical costs, focusing on the advantages and disadvantages of Random Forest and Gradient Boosting regressors. To improve interoperability with the models, it also incorporates data pretreatment methods like label encoding to convert categorical data into numerical representations. Third, it reduces overfitting

and enhances generalisation by using K-fold cross-validation to provide a reliable evaluation of model performance. Ultimately, it offers valuable insights that healthcare institutions can utilise to more effectively allocate resources, identify high-cost patients, and streamline budget planning.

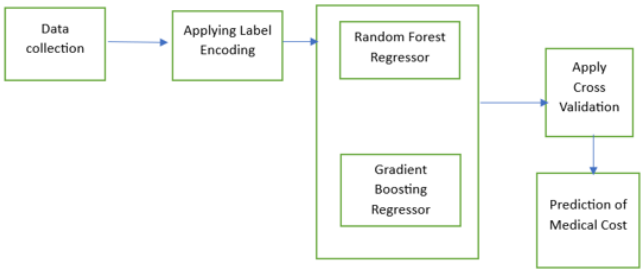
The study utilises a real-world dataset obtained from Kaggle, which includes a diverse range of patient profiles and medical histories, ensuring a comprehensive evaluation of cost prediction techniques. The comparative study of several machine learning techniques, such as Random Forest and Gradient Boosting regressors, for predicting medical costs also distinguishes it. The results are more stable and reliable thanks to the use of K-Fold cross-validation, which distinguishes this study from others. Additionally, by emphasising both technical performance and practical implications, the study bridges the gap between theoretical research and actual implementations, thereby enhancing patient care and reducing costs.

II. LITERATURE REVIEW

Interest in and research on creating precise models for predicting medical expenses has increased recently at the intersection of machine learning (ML) and healthcare analytics. The capacity to predict healthcare. Optimizing resource allocation, budget planning, and healthcare administration all depend on charging according to patient characteristics and treatment qualities [1]. To estimate healthcare expenditures, traditional regression-based methods have long been used. These methods look at the correlations between predictors including age, gender, BMI, and smoking status [2]. However, researchers are now investigating more advanced machine learning techniques due to the limits of linear models in capturing complex patterns and non-linear interactions in medical data [3]. Because of their capacity to manage high-dimensional data and the non-linear correlations present in healthcare datasets, ensemble learning techniques like random forests and gradient boosting machines (GBMs) have become more well-known [4]. When it comes to healthcare cost estimate challenges, these ensemble methods outperform single-model approaches by utilizing several weak learners to produce solid predictions [5]. Furthermore, the development of deep learning has transformed healthcare prediction analytics. Deep neural networks have demonstrated potential in identifying latent features and temporal relationships for medical cost prediction because of their hierarchical architecture and capacity to extract complex patterns from unprocessed data [6]. New approaches to individualized healthcare cost modeling have been made possible by the adaptation of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to analyze time-series data and medical pictures, respectively [7]. The ethical and legal issues about data privacy, equity, and openness are crucial when using machine learning to forecast healthcare costs [8]. Yabin Paul Thomas et al. represented a work on data mining approaches in healthcare. WEKA, ODND, NCC2, RST, ANN, SPSSModeler, and SNP are the different data mining tools analyzed by the authors [9]. Methods like feature selection, dimensionality reduction, and outlier detection are essential for enhancing prediction performance and enabling domain-specific insights [10]. In the future, combining cutting-edge machine learning techniques with domain knowledge and actual medical data has enormous potential to revolutionise healthcare delivery and reduce costs. Researchers and practitioners seek to improve patient outcomes, maximize resource allocation, and reduce the financial risks related to medical treatments by utilizing data-driven methodologies [11]. Furthermore, utilising natural language processing (NLP) in healthcare cost prediction models can yield valuable insights from electronic health records (EHRs) and unstructured clinical notes. By capturing subtle patient information that standard structured data could miss, natural language processing (NLP) approaches allow the extraction of relevant features from textual data, improving the prediction potential of machine learning (ML) models [12]. Furthermore, transfer learning has demonstrated potential in healthcare analytics by enhancing model performance and lowering the requirement for sizable labelled datasets. By utilizing existing knowledge, transfer learning enables models developed on massive, generic datasets to be optimized for healthcare applications, speeding up the creation of precise predictive models for healthcare costs [13]. Another frontier in predicting healthcare costs is the integration of genetic data with conventional clinical and demographic data. ML models can produce more accurate risk categorization and customized cost projections by integrating genetic data, allowing for more individualized treatment regimens and more efficient use of resources [14]. The creation of hybrid models, which blend machine learning approaches with conventional statistical methods, is another crucial area of research. By combining the best features of both approaches, these hybrid systems seek to increase the robustness and interpretability of healthcare cost prediction [15]. Furthermore, a promising field is the use of reinforcement learning (RL) in healthcare cost prediction. Cost prediction models that adapt to shifting patient situations and healthcare scenarios can be created using RL algorithms, which learn optimal policies through interactions with the environment [16].

III. PROPOSED MODEL

In this work, a machine learning-based model has been proposed for predicting medical costs using a Random Forest regressor. The block diagram of the proposed model is shown in Figure 1 below.



[Fig.1: Proposed Model for Medical Cost Prediction]

The proposed algorithm has been shown in table 1 below:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table I: Proposed Algorithm of Medical Cost Prediction Applying Different Machine Learning Approaches

Step 1:	Date set collection: $D_{medicalcost} = Data(X:Y)$
Step 2:	Applying Line Encoding for converting text data to numeric $For\ All\ (X = text\ value)\ \ applying\ Label\ Encoding$ $D_{medicalcost_{new}}(X_{numeric}) : f_{labelencoding}(D_{medicalcost}(X_{text}))$
Step 3:	Applying Machine learning models: Applying Random Forest Regressor with K-Fold: $RF = f_{randomforest}(D_{medicalcost_{new}})$ $K_{randomforest} = f_{KFold}(n_{split}, suffle, Ramdomstate)$ $C_{RF} = f_{crossvalidation}(RF, D_{medicalcost_{new}}(X_{numeric}), y, K_{randomforest})$ Applying Linear Regression: $LR = f_{linearregression}(D_{medicalcost_{new}})$ $K_{linearregression} = f_{KFold}(n_{split}, suffle, Ramdomstate)$ $C_{LR} = f_{crossvalidation}(LR, D_{medicalcost_{new}}(X_{numeric}), y, K_{linearregression})$) Applying Random Gradient Boosting Regressor: $GB = f_{gradientboosting}(D_{medicalcost_{new}})$ $K_{gradientboosting} = f_{KFold}(n_{split}, suffle, Ramdomstate)$ $C_{GB} = f_{crossvalidation}(RF, D_{medicalcost_{new}}(X_{numeric}), y, K_{gradientboosting})$
Step 4:	Model Validation: Calculating Root Mean Square Error of each model $RMSE_{RF} = f_{rmse}(C_{RF})$ $RMS_{LR} = f_{rmse}(C_{LR})$ $RMSE_{GB} = f_{rmse}(C_{GB})$

IV. DATASET DESCRIPTION

The dataset used for the experiments was collected from Kaggle, consisting of 1,338 patient records. The dataset comprises the following features.

- Age: Age of the person.
- Sex: Gender of the person (Male or Female).
- BMI: Body Mass Index, a measure of body fat based on height and weight.
- Children: Total number of children the person has.
- Smoker: Indicates whether the person is a smoker (Yes or No).
- Region: Geographic region where the person resides (Southwest, Southeast, Northeast,

Northwest).

The dataset comprises 1,338 records (rows) and seven attributes (columns). The target variable for this analysis is "charges," which represents the medical insurance cost for each individual.

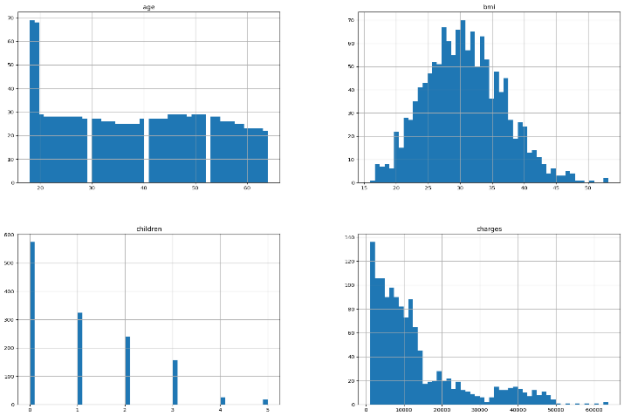
A snapshot of the dataset description is shown in Figure 2 below.

```
[ ] Med.describe()
```

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

[Fig.2: Dataset Description]

The distribution of age, BMI, children, and charges is shown in Figures 3, 4, and 5 below.



[Fig.3: Age, BMI, Children and Charges Distribution Over the Dataset

A. Applying Label Encoding

In the above-described dataset, the fields 'sex', 'smoker', and 'region' are text fields. Label Encoding has been applied to these fields to convert them into numeric categories. The updated field values are shown in Figure 3 below.

```
[ ] print(X,Y)
```

	age	sex	bmi	children	smoker	region
0	19	0	27.900	0	1	3
1	18	1	33.770	1	0	2
2	28	1	33.000	3	0	2
3	33	1	22.705	0	0	1
4	32	1	28.880	0	0	1
...	...	...	...	...	...	...
1333	50	1	30.970	3	0	1
1334	18	0	31.920	0	0	0
1335	18	0	36.850	0	0	2
1336	21	0	25.800	0	0	3
1337	61	0	29.070	0	1	1

[Fig.4: Dataset After Applying Label Encoding]

B. Applying Classifier

This study aims to forecast insurance prices based on

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

several variables, including age, sex, number of children, region, BMI, and smoking status. For predicting medical insurance costs, Random Forest Regressor and Linear Regressor models have been applied.

For predicting numerical values, Random Forest Regression is a flexible machine-learning method. It improves accuracy and lessens overfitting by combining the predictions of several decision trees.

By utilizing a related and known data value, the data analysis method known as linear regression can be used to forecast the value of unknown data. It utilizes a linear equation to quantitatively represent the relationship between the known or independent variable and the unknown or dependent variable.

C. Applying Cross-Validation Technique

Grid Search Cross-Validation involves selecting the best parameters for a model by evaluating different combinations using cross-validation. With K-Fold Cross-Validation, the dataset is divided into 'k' consecutive folds. The model is repeatedly trained on 'k-1' folds, and the remaining fold is used for validation. This method reduces overfitting and enhances reliability.

```
[ ] X,y=Med.drop("charges",axis=1),Med["charges"]
    rf = RandomForestRegressor(n_estimators=100, random_state=42)
    k = 5
    kf1 = KFold(n_splits=k, shuffle=True, random_state=42)
    mse_scores = cross_val_score(rf, X, y, scoring='neg_mean_squared_error', cv=kf11)
    rmse_scores = np.sqrt(-mse_scores)
    print("Cross-Validation Scores (RMSE):", rmse_scores)
    print("Mean RMSE:", np.mean(rmse_scores))

Cross-Validation Scores (RMSE): [4555.01535319 4919.37160168 4735.30451227 5003.9775205  5174.59852344]
Mean RMSE: 4877.653502213847
```

[Fig.5: Snapshot of Proposed Random Forest Regressor with Cross Validation Model]

```
lr = LinearRegression()
k = 5
kf2 = KFold(n_splits=k, shuffle=True, random_state=42)
mse_scores = cross_val_score(lr, X, y, scoring='neg_mean_squared_error', cv=kf2)
rmse_scores = np.sqrt(-mse_scores)
print("Cross-Validation Scores (RMSE):", rmse_scores)
print("Mean RMSE:", np.mean(rmse_scores))

Cross-Validation Scores (RMSE): [5799.58709144 6092.79444703 5782.5615577  6446.58476428 6232.10173577]
Mean RMSE: 6070.7259192420415
```

[Fig.6: Snapshot of Proposed Linear Regression with Cross Validation Model]

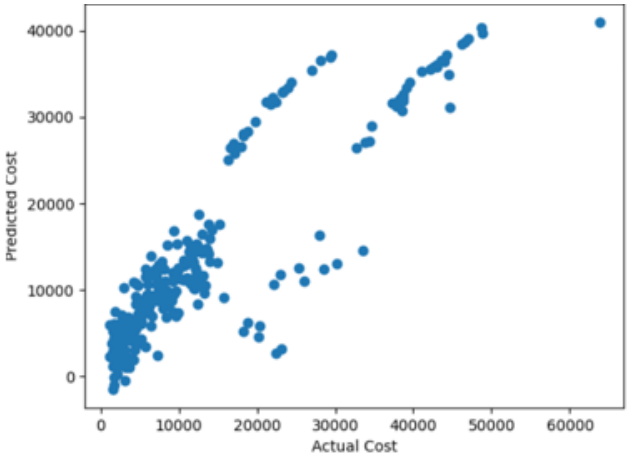
```
g=GradientBoostingRegressor(n_estimators=100, max_depth=7)
k = 5
kf = KFold(n_splits=k, shuffle=True, random_state=42)
mse_scores = cross_val_score(g, X, y, scoring='neg_mean_squared_error', cv=kf)
rmse_scores = np.sqrt(-mse_scores)
print("Cross-Validation Scores (RMSE):", rmse_scores)
print("Mean RMSE:", np.mean(rmse_scores))

Cross-Validation Scores (RMSE): [4916.6498263  5297.21151962 5198.29947964 5314.71335425 5392.63144648]
Mean RMSE: 5223.901125257631
```

[Fig.7: Snapshot of Proposed Gradient Boosting Regressor with Cross Validation Model]

V. RESULT ANALYSIS AND DISCUSSION

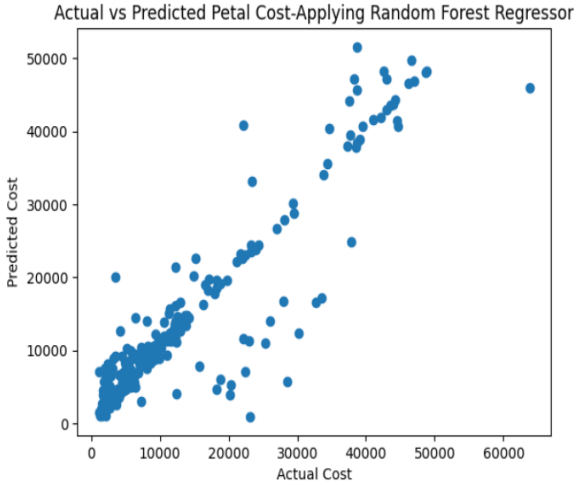
Three key metrics are used to evaluate the accuracy of prediction models: Mean Absolute Error (MAE), R-squared score, and Root Mean Squared Error (RMSE). By taking the square root of the average squared deviations between the actual and projected values, or RMSE, one can determine how well the model predicts by assigning more weight to higher errors. A value closer to 1 implies a better fit. The R^2 score, also known as the coefficient of determination, indicates the percentage of variance in the dependent variable that the independent variables can explain. Without considering the direction of the errors, MAE computes the average of the absolute disparities between actual and anticipated values to provide a clear indicator of the average error in the model's predictions. When combined, these measures provide a thorough understanding.

The Mean Absolute Error, R-squared score, and Root Mean Squared Error score achieved by applying Linear Regression are 4,154.1223, 0.7855, and 6,171.7245, respectively. The performance of the linear regression model is shown in Figure 8 below, which displays the actual cost versus the predicted cost graph.



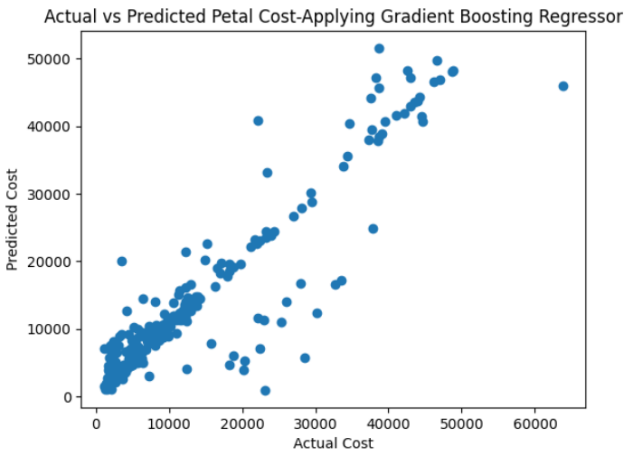
[Fig.8: Predicted Cost Applying Linear Regression]

Mean Absolute Error, R^2 Score, and Root Mean Squared Error scores achieved applying Random Forest Regressor are 2832.2254, 0.8445, 4871.6588. The performance of the Random Forest Regressor model is shown in Figure 9 below, which displays the actual cost versus the predicted cost graph.



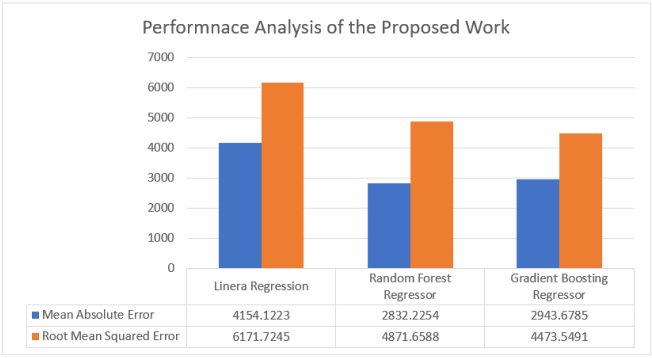
[Fig.9: Predicted Cost Applying Random Forest Regressor]

The Mean Absolute Error, R-squared score, and Root Mean Squared Error score achieved by applying the Gradient Boosting Regressor are 2943.6785, 0.8567, and 4473.5491, respectively. The performance of the Gradient Boosting Regressor model is illustrated in Figure 10 below, which compares the actual cost with the predicted cost graph.

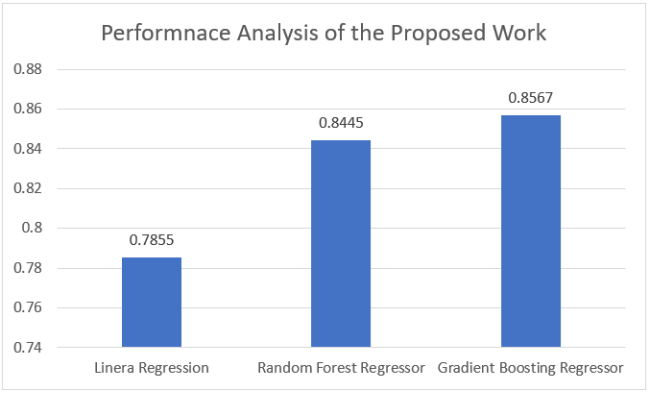


[Fig.10: Predicted Cost Applying Gradient Regressor]

A summary of the proposed work's performance is shown in Figures 11 and 12 below.



[Fig.11: Performance Analysis of the Proposed Work Representing Mean Absolute Error and Root Mean Squared Error]



[Fig.12: Performance Analysis of the Proposed Work Representing R^2 Score]

The performance metrics of three regression models—Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor—are compared to shed light on how well each model predicts medical costs. The results show that the Gradient Boosting Regressor performs the best, obtaining the most outstanding R^2 score (0.8567), which indicates that it can account for the majority of the variance in the data. In comparison to the other models, it also has the lowest Root Mean Squared Error (RMSE) (4473.5491), indicating more accurate predictions. In terms of dependability in predicting medical costs, Random Forest Regressor comes in close second with a high R^2 score (0.8445) and lower RMSE (4871.6588) and Mean Absolute Error (MAE) (2832.2254). Linear regression, on the other hand, has a respectable R2 (0.7855) but is the least effective of the three models due to its larger RMSE (6171.7245) and MAE (4154.1223), which indicate that it is less accurate. Our findings suggest that more complicated models, such as Random Forest and Gradient Boosting, can more accurately and consistently predict medical costs by better capturing their complexity.

Comparative analysis of this work with a few existing works in this domain has been presented in Table 1 below.

Proposed Work	Observations
[2]	Applied Random Forest for cost prediction. It has been demonstrated that the causal forest method achieved better outcomes than classical methods.
[3]	By using expenses as a stand-in for health, a healthcare algorithm exhibits racial bias, resulting in fewer Black patients obtaining necessary care even when they are sicker. More Black patients could receive the assistance they require if this bias were changed.
[6]	Applied a Deep learning approach for health care data storage. Average AUCROC score achieved: 0.93–0.94
[7]	The paper highlights issues, including interpretability and the need for improved techniques, while examining the potential of deep learning in healthcare to analyse complex data.
[8]	Applied locally supervised metric learning, based on a dataset of 15,000 patients, shows that personalised logistic regression models outperform global models in risk prediction using electronic health information.
[10]	Applied a regression-based Gaussian model to multivariate longitudinal clinical data.
[15]	Explores the possibilities and constraints of machine learning in the medical field, emphasizing the necessity of precise forecasts and the value of fusing AI with human knowledge to enhance patient care.
Our proposed model	Linear Regression, Random Forest, and Gradient Boosting have been applied to predict medical costs. Best performance achieved through Gradient Boosting. R^2 score (0.8567) (RMSE) (4473.5491)

VI. CONCLUSION

This research illustrated the efficacy of machine learning methodologies in predicting health insurance expenses based on individual demographic and lifestyle attributes. We made sure that the comparison of algorithms was fair and strong by using preprocessing steps like feature scaling and categorical encoding and checking models with k-fold cross-validation. When we tested Linear Regression, Random Forest, and Gradient Boosting, Gradient Boosting consistently outperformed the others in terms of R^2, MAE, and RMSE. This means that Gradient Boosting is particularly effective at identifying the complex, non-linear relationships between personal traits and insurance costs. These results show that advanced regression models could improve cost prediction in healthcare insurance analytics, which would lead to better pricing strategies and more personalised risk assessment. This study demonstrates that machine learning models, particularly Gradient Boosting, can be effective in predicting health insurance costs; however, certain limitations exist. The dataset utilised comprised only a restricted array of demographic characteristics, omitting essential elements such as medical history, income, and lifestyle habits, which could

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

influence prediction accuracy. The model's generalisability is also limited by the fact that the data is static and the population may not be very diverse. Advanced models are also hard to understand and complex, which makes them hard to use in sensitive areas like healthcare. For future work, incorporating more comprehensive and long-term datasets, exploring deep learning methods, and utilising explainable AI methods can enhance both accuracy and transparency. Furthermore, validating models across varied populations, creating real-time predictive tools, and tackling ethical concerns will be essential for enhancing the practical utilisation of these models in health insurance analytics.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. The New England Journal of Medicine, 375(13), 1216-1219. DOI: https://doi.org/10.1056/nejmp1606181
2. Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association, 113(523), 1228-1242. DOI: https://doi.org/10.1080/01621459.2017.1319839
3. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447- 453. DOI: https://doi.org/10.1126/science.aax2342
4. Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 24(1), 198-208. DOI: https://doi.org/10.1093/jamia/ocw042
5. Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370. DOI: https://doi.org/10.1093/jamia/ocw112
6. Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning for electronic health records. npj Digital Medicine, 1, 18. DOI: https://doi.org/10.1038/s41746-018-0029-1
7. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep Learning for Healthcare: A Review, Opportunities, and Challenges. Briefings in Bioinformatics, 19(6), 1236-1246. DOI: https://doi.org/10.1093/bib/bbx044
8. Ng, K., Sun, J., Hu, J., Wang, F., & Shen, Y. (2017). Personalized predictive modeling and risk factor identification using patient similarity. AMIA Annual Symposium Proceedings, 2015, 1176-1185. https://pubmed.ncbi.nlm.nih.gov/26306255/
9. Paul Thomas, Yabin. (2024). Application Of Data Mining In Health Care. International Research Journal of Modernisation in Engineering, Technology, and Science. 06. 2582-5208. DOI: https://www.doi.org/10.56726/IRJMETS7375510
10. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2017). Predicting disease progression with a model combining sequence and non-sequence data. International Conference on Machine Learning (ICML). https://proceedings.mlr.press/v56/Futoma16.html
11. Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. JAMA, 322(18), 1806- 1816. DOI: https://doi.org/10.1001/jama.2019.16489
12. Davenport, T., & Kalakota, R. (2019). The Potential for Artificial Intelligence in Healthcare Future Healthcare Journal, 6(2), 94-98. DOI: https://doi.org/10.7861/futurehosp.6-2-94
13. Shah, N. D., Steyerberg, E. W., & Kent, D. M. (2018). Big Data and Predictive Analytics: Recalibrating Expectations. Journal of the American Medical Association, 320(1), 27-28. DOI: https://doi.org/10.1001/jama.2018.5602
14. Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. JAMA, 319(13), 1317-1318. DOI: https://doi.org/10.1001/jama.2017.18391
15. Chen, J. H., & Asch, S. M. (2017). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. The New England Journal of Medicine, 376(26), 2507-2509. DOI: https://doi.org/10.1056/nejmp1702071
16. Rutter, J. L., & Boudreault, D. J. (2019). Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Approaches. Applied Clinical Informatics, 10(5), 844-846. DOI: https://doi.org/10.3346/jkms.2020.35.e379

AUTHOR'S PROFILE

Dr. Annwesha Banerjee Majumder is working as an Asst. Professor in the Department of Information Technology, JIS College of Engineering. She completed her B.Tech in Information Technology from JIS College of Engineering in 2006. She completed her master's degree in Mobile Communication and Network Technology in 2009. She achieved a doctorate in Computer Science and Engineering from Maulana Abul Kalam Azad University of Technology. She has more than 15 years of teaching experience and 10 years of research experience. She is a live member of FOSET. She has published over 20 research papers and holds 15 patents. Her research interests include machine learning, Data Analytics, and Cloud Computing.

Dr Sumit Das is currently working as an Associate Professor in the Department of Information Technology at JIS College of Engineering, an Autonomous Institution under the JIS Group, India. He obtained both B.Tech (IT) in 2006, M.Tech (CSE) in 2008 and PhD (ETM-KU) In 2021, from the University of Kalyani. His areas of interest include Artificial Intelligence (AI), Machine Learning (ML), Data Science, IT & Organisation, and Soft Computing in the fields of medical diagnosis and healthcare management. He is pursuing an executive MBA from the Indian Institute of Technology Roorkee. He has published more than 40 indexed research papers in reputable international conferences and journals.

Aniruddha Biswas, University Research Scholar of the Department of Biochemistry & Biophysics, University of Kalyani, Kalyani, Nadia, West Bengal. He has completed his M.E. in Software Engineering from the Department of Information Technology at Jadavpur University, Salt Lake Campus, Kolkata, West Bengal, and his B.Tech. from DETS (formerly USIC) at the University of Kalyani, Kalyani, Nadia, West Bengal. He qualified for the GATE in 2004 with a percentile of 95.02 in Information Technology, ranking 462nd all-India. Worked as a Software Engineer in HCL Technologies Pvt. Ltd., Noida, UP, India, for approximately 5 Years for client IKEA and currently working as an Assistant Professor in the Department of Information Technology, JIS College of Engineering, Kalyani, Nadia, West Bengal, since 2012 to date. He has published papers in various journals and conferences of international repute in multiple domains. His research interests are in data science, bioinformatics (LNCRNA and human diseases). He has a publication in Springer Nature in Applied

Biochemistry and Biotechnology, 2021.

Trishita Ghosh is currently serving as an Assistant Professor in the Department of Information Technology at Guru Nanak Institute of Technology (GNIT), India. She holds an M. Tech degree from the University of Calcutta and a B.E. degree from UIT. With a strong academic background, she has been actively involved in both teaching and research. Her core research interests include machine learning, pattern recognition, image processing, and data science. Ms Ghosh is particularly focused on applying computational intelligence and data-driven methodologies to address and solve complex, real-world problems, contributing to the advancement of technology and innovation in her field.

Raj Poddar holds a Bachelor's degree in Computer Science and a Master's degree in Computer Applications. As a student at JIS College of Engineering and Kanchrapara College, he developed e-commerce web apps using HTML, CSS, JavaScript, PHP, and GitHub. He built several projects from scratch, including a Dairy Management System, a Food Delivery System, and an Amazon clone. He also worked on several projects, including insurance cost optimisation and e-commerce delivery time optimisation, utilising AI and ML techniques and algorithms, such as random forest. He also attended workshops on AWS Fundamentals, Web Development, and Cybersecurity Fundamentals. He is passionate about learning new technologies and applying them to real-world problems. He has completed my Master of Computer Applications degree, with a focus on computer programming and specific applications.

Suchetana Chakraborty is currently pursuing my final year of B. Tech in Information Technology from Guru Nanak Institute of Technology. Her academic interests include Machine Learning and Deep Learning, with a keen focus on their applications in real-world problem solving. She has worked on projects involving online handwritten Bangla character recognition and ATM Recognition. She has also participated in several technical fests and exhibitions, showcasing innovative models and creative solutions. Passionate about research and continuous learning, she aims to contribute to advancements in Artificial Intelligence while pursuing higher studies and impactful research in the field of intelligent computing. Her research interests are Artificial Intelligence.