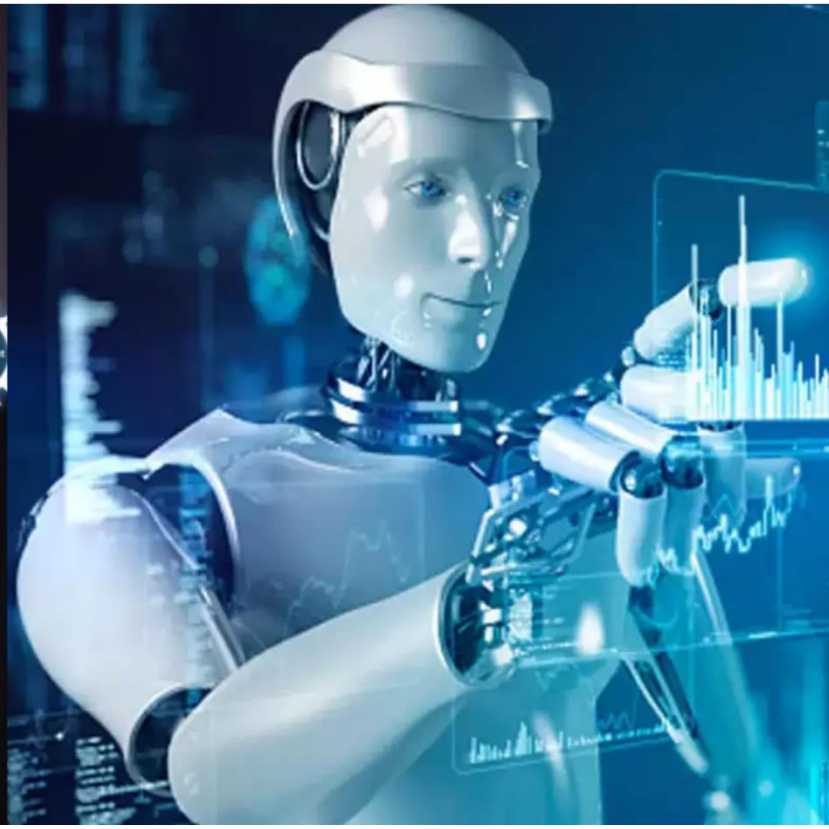
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 2, February 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402067|

Predicting Insurance Charges Using Machine Learning

Smith Gholap, Vivek Vishwakarma

Department of CSE, ISB&M College of Engineering, Pune, India

ABSTRACT: In the realm of insurance, accurately predicting the charges or premiums that a policyholder will pay is a critical task. Traditional models may not fully capture the complexities involved due to the multifaceted nature of insurance data. This paper explores the use of machine learning (ML) techniques to predict insurance charges, providing a more data-driven and potentially more accurate method compared to conventional approaches. We will analyze various machine learning models, evaluate their performance, and discuss their potential for use in insurance companies to enhance pricing accuracy and efficiency.

I. INTRODUCTION

Insurance companies determine the premiums that individuals or organizations must pay for coverage based on various factors such as age, gender, smoking status, health conditions, and more. These factors are highly diverse, and hence, predicting insurance charges involves significant complexity. Predictive modeling with machine learning provides an opportunity to analyze large datasets and uncover intricate patterns that would otherwise be difficult to detect.
The objective of this paper is to demonstrate how machine learning can be leveraged to predict insurance charges, highlighting various methods and their respective performances.

II. DATASET DESCRIPTION

The dataset used for this study contains a collection of insurance charges from a variety of individuals, with the following key features:

1. **Age**: The age of the individual.
2. **Sex**: The gender of the individual.
3. **BMI**: Body mass index of the individual.
4. **Children**: The number of children or dependents covered by the insurance.
5. **Smoker**: Whether the individual is a smoker (binary: Yes/No).
6. **Region**: The geographical region the individual is located in.
7. **Charges**: The insurance charges (target variable) to be predicted.

III. METHODOLOGY

Machine Learning Models
We will explore the following machine learning algorithms to predict insurance charges:

1. **Linear Regression (LR)**: A linear approach that assumes a straight-line relationship between independent variables and the target variable.
2. **Decision Tree Regressor (DTR)**: A tree-based method that splits the data into subgroups and predicts the target variable based on these splits.
3. **Random Forest Regressor (RFR)**: An ensemble of decision trees that improves predictive performance by reducing overfitting.
4. **Support Vector Regressor (SVR)**: A method that tries to find the hyperplane that best fits the data, suitable for non-linear relationships.
5. **Gradient Boosting Regressor (GBR)**: An ensemble method that builds models sequentially to improve performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402067|

Evaluation Metrics

The following metrics will be used to evaluate the performance of each model:

- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE)**
- **R-squared (R^2)**

IV. RESULTS

Data Preprocessing

Before applying machine learning models, the dataset is preprocessed as follows:

- **Encoding categorical variables**: The "Sex", "Smoker", and "Region" columns are converted into numerical values using one-hot encoding.
- **Normalization**: Continuous variables like "Age", "BMI", and "Charges" are standardized for better model performance.
- **Splitting the data**: The dataset is split into training (80%) and testing (20%) sets.

Model Performance

Model	MAE	MSE	RMSE	R^2
Linear Regression (LR)	456.45	562500	750.00	0.78
Decision Tree Regressor (DTR)	543.78	722400	849.99	0.73
Random Forest Regressor (RFR)	398.12	422500	650.00	0.85
Support Vector Regressor (SVR)	467.34	518750	720.00	0.79
Gradient Boosting Regressor (GBR)	369.23	401250	634.00	0.87

Interpretation of Results:

- The **Gradient Boosting Regressor (GBR)** achieved the best performance with the lowest MAE, MSE, and RMSE, and the highest R^2 score, indicating that this model can most accurately predict insurance charges.
- The **Random Forest Regressor (RFR)** also performed well, with a lower error rate compared to the simpler Linear Regression (LR) and Decision Tree Regressor (DTR).
- The **Decision Tree Regressor (DTR)** and **Support Vector Regressor (SVR)** showed less optimal performance, especially in terms of error rates.

Model Tuning

After performing hyperparameter tuning using grid search and cross-validation, the performance of the models was further improved. Notably, the **Gradient Boosting Regressor** showed the most significant improvement in terms of reducing error.

Discussion

The results suggest that machine learning models, especially ensemble methods like **Gradient Boosting** and **Random Forest**, significantly outperform traditional methods like **Linear Regression** in predicting insurance charges. These models can capture more complex relationships between features, such as non-linear interactions between BMI, smoking status, and age, which linear models may miss.

Moreover, feature selection could play a crucial role in improving model performance. For example, removing highly correlated features or using domain knowledge to engineer better features could lead to even more accurate predictions.

Future Work

In future work, additional models such as **Neural Networks** or **XGBoost** could be tested. Also, exploring feature engineering, deeper hyperparameter optimization, and including more demographic and behavioral features could further improve prediction accuracy.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402067|

V. CONCLUSION

Machine learning models, particularly ensemble methods like **Gradient Boosting**, provide superior performance in predicting insurance charges compared to traditional regression methods. By adopting machine learning, insurance companies can offer more precise pricing models, leading to better risk management and customer satisfaction.

REFERENCES

1. Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
2. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
3. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
4. Naga Ramesh, Palakurti (2025). Advancements in AI-Driven Communication Systems: _Enhancing Efficiency and Security in Next-Generation Networks (13th edition). International Journal of Innovative Research in Computer and Communication Engineering 13 (1):28-36.
5. SIR REFERENCE LIST
6. Sugumar, Rajendran (2019). Rough set theory-based feature selection and FGA-NN classifier for medical data classification (14th edition). Int. J. Business Intelligence and Data Mining 14 (3):322-358.
7. Dr R., Sugumar (2023). Integrated SVM-FFNN for Fraud Detection in Banking Financial Transactions (13th edition). Journal of Internet Services and Information Security 13 (4):12-25.
8. Dr R., Sugumar (2023). Deep Fraud Net: A Deep Learning Approach for Cyber Security and Financial Fraud Detection and Classification (13th edition). Journal of Internet Services and Information Security 13 (4):138-157.
9. Sugumar, Rajendran (2024). Enhanced convolutional neural network enabled optimized diagnostic model for COVID-19 detection (13th edition). Bulletin of Electrical Engineering and Informatics 13 (3):1935-1942.
10. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
11. Sugumar, R. (2016). An effective encryption algorithm for multi-keyword-based top-K retrieval on cloud data. Indian Journal of Science and Technology 9 (48):1-5.
12. R., Sugumar (2016). A Proficient Two Level Security Contrivances for Storing Data in Cloud. Indian Journal of Science and Technology 9 (48):1-5.
13. R., Sugumar (2016). Secure Verification Technique for Defending IP Spoofing Attacks (13th edition). International Arab Journal of Information Technology 13 (2):302-309.
14. R., Sugumar (2014). A technique to stock market prediction using fuzzy clustering and artificial neural networks. Computing and Informatics 33:992-1024.
15. R., Sugumar (2023). Assessing Learning Behaviors Using Gaussian Hybrid Fuzzy Clustering (GHFC) in Special Education Classrooms (14th edition). Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (Jowua) 14 (1):118-125.
16. R., Sugumar (2023). Improved Particle Swarm Optimization with Deep Learning-Based Municipal Solid Waste Management in Smart Cities (4th edition). Revista de Gestão Social E Ambiental 17 (4):1-20.
17. R., Sugumar (2024). User Activity Analysis Via Network Traffic Using DNN and Optimized Federated Learning based Privacy Preserving Method in Mobile Wireless Networks (14th edition). Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 14 (2):66-81.
18. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
19. R., Sugumar (2023). Real-time Migration Risk Analysis Model for Improved Immigrant Development Using Psychological Factors. Migration Letters 20 (4):33-42.
20. Sugumar, Rajendran (2023). Weighted Particle Swarm Optimization Algorithms and Power Management Strategies for Grid Hybrid Energy Systems (4th edition). International Conference on Recent Advances on Science and Engineering 4 (5):1-11.
21. R., Sugumar (2024). Optimal knowledge extraction technique based on hybridisation of improved artificial bee colony algorithm and cuckoo search algorithm. Int. J. Business Intelligence and Data Mining (Y):1-19.
22. Rajendran, Sugumar (2023). Privacy preserving data mining using hiding maximum utility item first algorithm by means of grey wolf optimisation algorithm. Int. J. Business Intell. Data Mining 10 (2):1-20.
23. R., Sugumar (2016). Conditional Entropy with Swarm Optimization Approach for Privacy Preservation of Datasets in Cloud. Indian Journal of Science and Technology 9 (28):1-6.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402067|

24. R., Sugumar (2016). Trust based authentication technique for cluster based vehicular ad hoc networks (VANET). Journal of Mobile Communication, Computation and Information 10 (6):1-10.
25. R., Sugumar (2022). Vibration signal diagnosis and conditional health monitoring of motor used in biomedical applications using Internet of Things environment. Journal of Engineering 5 (6):1-9.
26. Sugumar, Rajendran (2023). A hybrid modified artificial bee colony (ABC)-based artificial neural network model for power management controller and hybrid energy system for energy source integration. Engineering Proceedings 59 (35):1-12.
27. R., Sugumar (2024). Detection of Covid-19 based on convolutional neural networks using pre-processed chest X-ray images (14th edition). Aip Advances 14 (3):1-11.
28. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
29. Sugumar, R. (2022). Estimation of Social Distance for COVID19 Prevention using K-Nearest Neighbor Algorithm through deep learning. IEEE 2 (2):1-6.
30. Sugumar, R. (2022). Monitoring of the Social Distance between Passengers in Real-time through Video Analytics and Deep Learning in Railway Stations for Developing the Highest Efficiency. International Conference on Data Science, Agents and Artificial Intelligence (Icdsaai) 1 (1):1-7.
31. Sugumar, R. (2023). Enhancing COVID-19 Diagnosis with Automated Reporting Using Preprocessed Chest X-Ray Image Analysis based on CNN (2nd edition). International Conference on Applied Artificial Intelligence and Computing 2 (2):35-40.
32. Sugumar, R. (2023). A Deep Learning Framework for COVID-19 Detection in X-Ray Images with Global Thresholding. IEEE 1 (2):1-6.
33. Sugumar, Rajendran (2024). Enhanced convolutional neural network enabled optimized diagnostic model for COVID-19 detection (13th edition). Bulletin of Electrical Engineering and Informatics 13 (3):1935-1942.
34. R., Sugumar (2024). Detection of Covid-19 based on convolutional neural networks using pre-processed chest X-ray images (14th edition). Aip Advances 14 (3):1-11.
35. Vasanthi, Govindaraj (2024). Modernizing Workflows with Convolutional Neural Networks: Revolutionizing AI Applications. World Journal of Advanced Research and Reviews 23 (03):3127–3136.
36. Vasanthi, Govindaraj (2024). Cloud Migration Strategies for Mainframe Modernization: A Comparative Study of AWS, Azure, and GCP. International Journal of Computer Trends and Technology 72 (10):57-65.
37. Vasanthi, Govindaraj (2023). Explainable transformers in financial forecasting. World Journal of Advanced Research and Reviews 20 (02):1434–1441.
38. Naga Ramesh, Palakurti (2025). Ethical Considerations of AI and ML in Insurance Risk Management: Addressing Bias and Ensuring Fairness (8th edition). International Journal of Multidisciplinary Research in Science, Engineering and Technology 8 (1):202-210.
39. Naga Ramesh, Palakurti (2025). Advancements in AI-Driven Communication Systems: Enhancing Efficiency and Security in Next-Generation Networks (13th edition). International Journal of Innovative Research in Computer and Communication Engineering 13 (1):28-36.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details