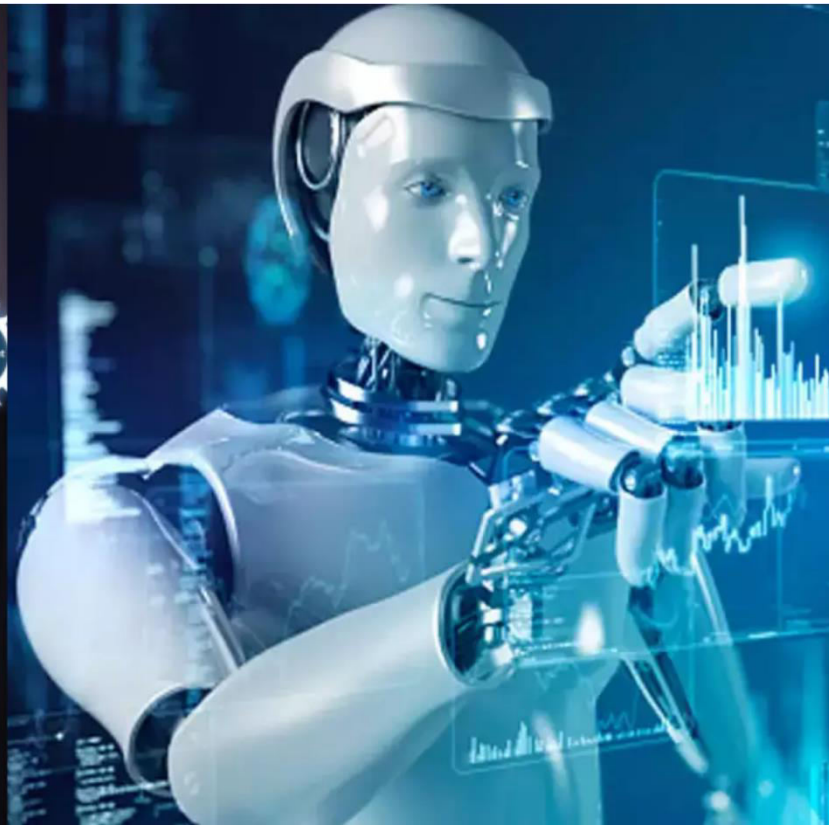
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 12, December 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 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412086|

Data Analysis and Prediction for Health Insurance Claim Approval and Risk Assessment

Anjan G[1], Naseerhusen Ankalagi[2]

PG Student, Dept. of MCA, City Engineering College, Bengaluru, India[1]

Assistant Professor, Dept. of MCA, City Engineering College, Bengaluru, India[2]

ABSTRACT: The rapid expansion of digital healthcare services has resulted in a significant increase in health insurance claim records, making conventional claim evaluation processes inefficient and inconsistent. Manual verification methods are time-consuming and often fail to detect complex risk patterns. This project proposes an intelligent data-driven framework for predicting health insurance claim approval and assessing claim risk using machine learning techniques. Historical insurance data is analyzed to extract meaningful insights related to policyholder attributes, medical details, treatment information, and claim characteristics. Data preprocessing steps such as cleaning, encoding, normalization, and feature selection are applied to enhance data quality. Several classification models are implemented and evaluated to determine claim approval likelihood and associated risk levels. According to experimental findings, the suggested method increases prediction accuracy, supports early identification of high-risk claims, and enhances decision-making efficiency in insurance claim management.

I. INTRODUCTION

Health insurance plays a vital part in protecting individuals from unexpected medical expenses while ensuring financial stability for healthcare services. With the growing number of insured individuals and as healthcare expenses rise, insurance companies must handle a large volume of claims on a regular basis. Traditional claim processing systems rely heavily on predefined rules and manual inspection, which often results in delayed settlements, higher operational costs, and inconsistent decisions.

Advancements in data analytics and machine learning have created opportunities to improve claim evaluation by leveraging historical data. Analyzing past claim records enables insurers to identify trends, assess claim validity, and understand factors influencing approval decisions. Predictive models can assist insurers in automating claim evaluation, reducing human bias, and improving transparency.

The goal of this project is to create a machine learning-based system that predicts health insurance claim approval and evaluates claim risk. By analyzing policyholder information, medical history, treatment details, and claim values, the system supports accurate and timely claim decisions. The proposed solution aims to enhance efficiency, reduce fraudulent actions, and ensure fair claim settlements through data-driven intelligence.

II. LITERATURE SURVEY

Numerous investigations have examined the application of data analytics as well as machine learning in insurance claim analysis. Earlier research emphasized the role of demographic and policy-related attributes in influencing claim approval decisions, highlighting the importance of historical data analysis. Further research revealed that machine learning classifiers outperform traditional rule-based systems by improving prediction accuracy and reducing manual effort.

Researchers have also investigated feature selection techniques to identify influential medical and policy variables, concluding that optimized feature sets enhance model performance and reduce computational complexity. Ensemble learning approaches, particularly tree-based models, have shown strong robustness when handling large and imbalanced insurance datasets. Recent works introduced risk categorization models that classify claims into multiple risk levels, enabling insurers to prioritize investigations and allocate resources efficiently. Overall, existing literature

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 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412086|

confirms that predictive analytics significantly improves claim processing precision and operational effectiveness in the health insurance domain.

III. METHODOLOGY

The suggested approach involves analyzing historical health insurance claim data to develop predictive models for claim approval and risk assessment. Data is where the process starts. collection from structured insurance datasets containing demographic details, medical information, policy attributes, claim amounts, and claim outcomes.

Preprocessing data is done in order to improve data quality and reliability. Missing values are addressed using suitable statistical methods, duplicate entries are removed, and inconsistent records are corrected Categorical variables are converted into numerical representations using encoding techniques, while numerical attributes are normalized to ensure balanced learning.

Exploratory data analysis is conducted to identify trends and relationships between input features and claim outcomes. Important features influencing approval and risk are chosen to train the model. A number of machine learning algorithms are then trained and evaluated using performance metrics to identify the most effective predictive model.

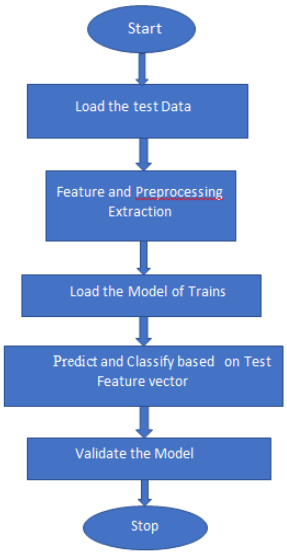


Fig.1: Activity Diagram

IV. SYSTEM DESIGN

The system design defines the overall structure and functional flow of the proposed health insurance claim approval and a system for evaluating risks. The system's purpose is to automate claim evaluation through data integration processing, predictive analytics, and decision support in a structured manner. The system begins with the data input module, where health insurance claim data is collected from authorized sources such as insurance databases or uploaded files. This data includes policyholder details, medical information, treatment records, policy coverage, and historical claim status. All incoming data is stored in a centralized database to ensure consistency, security, and easy retrieval. The preprocessing module is responsible for improving data quality before analysis. It eliminates duplicates and handles missing values. and inconsistent records, and converts categorical characteristics like policy type, disease category, and hospital type into numerical form using encoding techniques. Numerical attributes are normalized to ensure balanced learning by models for machine learning

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 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412086|

V. SYSTEM ARCHITECTURE AND DESIGN

The system's design as a modular framework that integrates data processing, predictive modeling, and decision support components. Claim data is collected from authorized sources and stored in a centralized database for consistency and accessibility. The preprocessing module cleans and prepares the data before forwarding it to the feature engineering layer.

In the machine learning layer, trained classification models analyze processed data to predict claim approval status and assess risk levels. The output of the system provides decision support by indicating whether a claim is likely to be approved, rejected, or requires further investigation. This structured architecture ensures scalability, accuracy, and ease of integration with existing insurance systems.

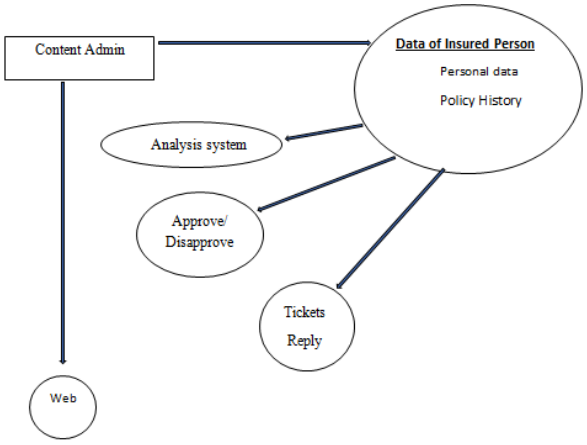


Fig 2: Flow Diagram

VI. IMPLEMENTATION

The implementation phase involved importing the health insurance claim dataset into the system and examining its structure and attributes. Preprocessing techniques methods were used to address missing values, eliminate duplicate records, and transform numerical variables from category one's format. Feature scaling was accustomed to maintain consistency across numerical attributes.

Training and testing were separated from the cleaned dataset. subsets. A To find patterns related to risk and claim approval, machine learning methods such as Random Forest, Decision Tree, and Logistic Regression were used. Metrics such as recall, accuracy, and precision, F1-score, and ROC-AUC were used to evaluate the model's effectiveness. In order to facilitate effective and trustworthy claim evaluation, The model that has been trained forecasts fresh claim data.

VII. RESULTS AND DISCUSSION

The proposed system was evaluated using real-world health insurance claim data. Among the implemented models, ensemble-based approaches demonstrated superior performance because of their capacity to catch complex feature interactions and reduce overfitting. Key factors such as claim amount, policy duration, previous claim history, and treatment type were found to significantly influence claim approval and risk classification.

The findings show that the predictive model effectively distinguishes between low-risk and high-risk claims, minimizing false approvals and supporting better financial Control. The results validate that machine learning-based analytics can significantly The precision and dependability of insurance claim decision-making.

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 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412086|

VIII. CONCLUSION

This project presented a data-driven framework for predicting health insurance claim approval and assessing claim risk using machine learning techniques. By analyzing historical claim data and identifying influential features, The suggested system improves decision accuracy and processing efficiency. The incorporation of predictive analytics reduces manual effort, supports early detection of high-risk claims, and enhances transparency in claim administration. The study shows how machine learning may be used to strengthen operational performance in the health insurance sector. Future enhancements may include real-time data integration and advanced learning models to further improve system reliability and scalability.

REFERENCES

1. Tom M. Mitchell, Machine Learning, McGraw-Hill, 1997.
2. Ian H. Witten, Eibe Frank, and Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.
3. Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, Elsevier, 2012.
4. Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, 2019.
5. Hastie, T., Tibshirani, R., and Friedman, J., The Elements of Statistical Learning, Springer, 2009.
6. Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E., "Machine Learning: A Review of Classification and Combining Techniques," Artificial Intelligence Review, 2006.
7. OECD, "Using Big Data Analytics in the Insurance Industry," OECD Publishing, 2017.
8. Rajkomar, A., Dean, J., and Kohane, I., "Machine Learning in Medicine," New England Journal of Medicine, 2019.
9. Brown, I., and Mues, C., "Comparing Classification Algorithms Experimentally for Unbalanced Credit Scoring Data Sets," Expert Systems with Applications, 2012.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

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