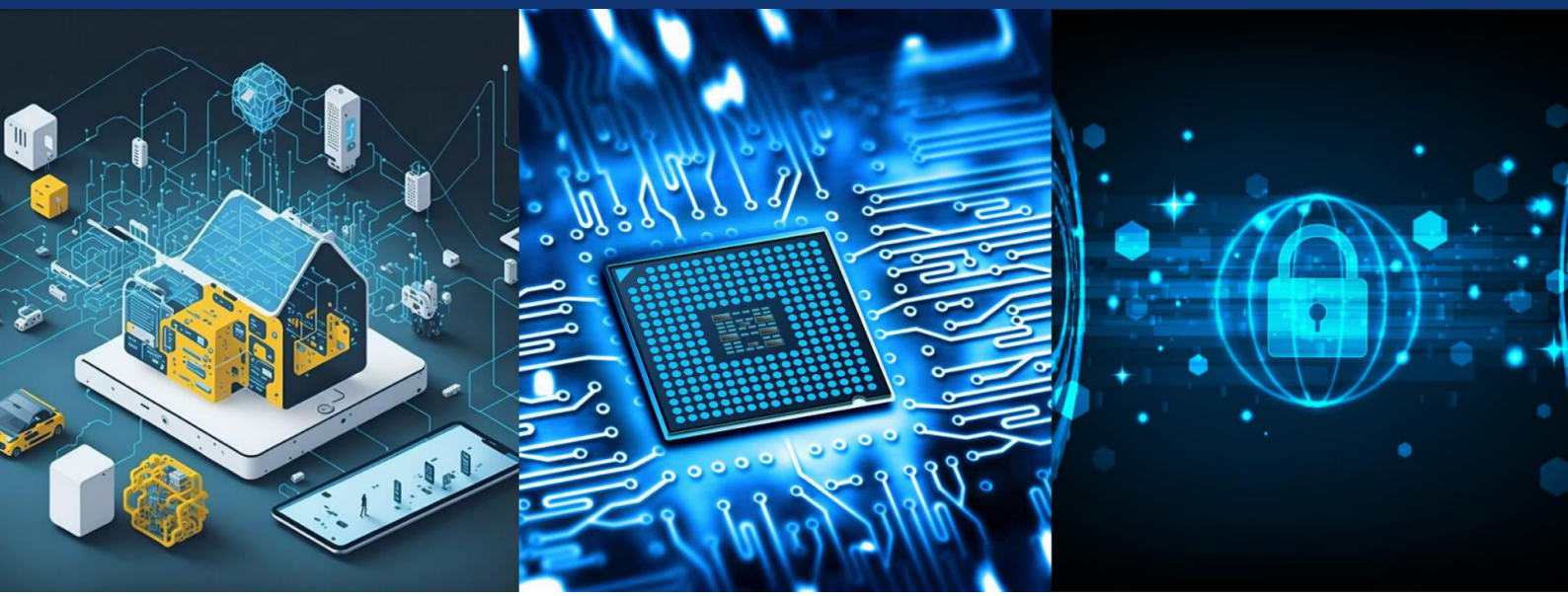
IJIRCCE

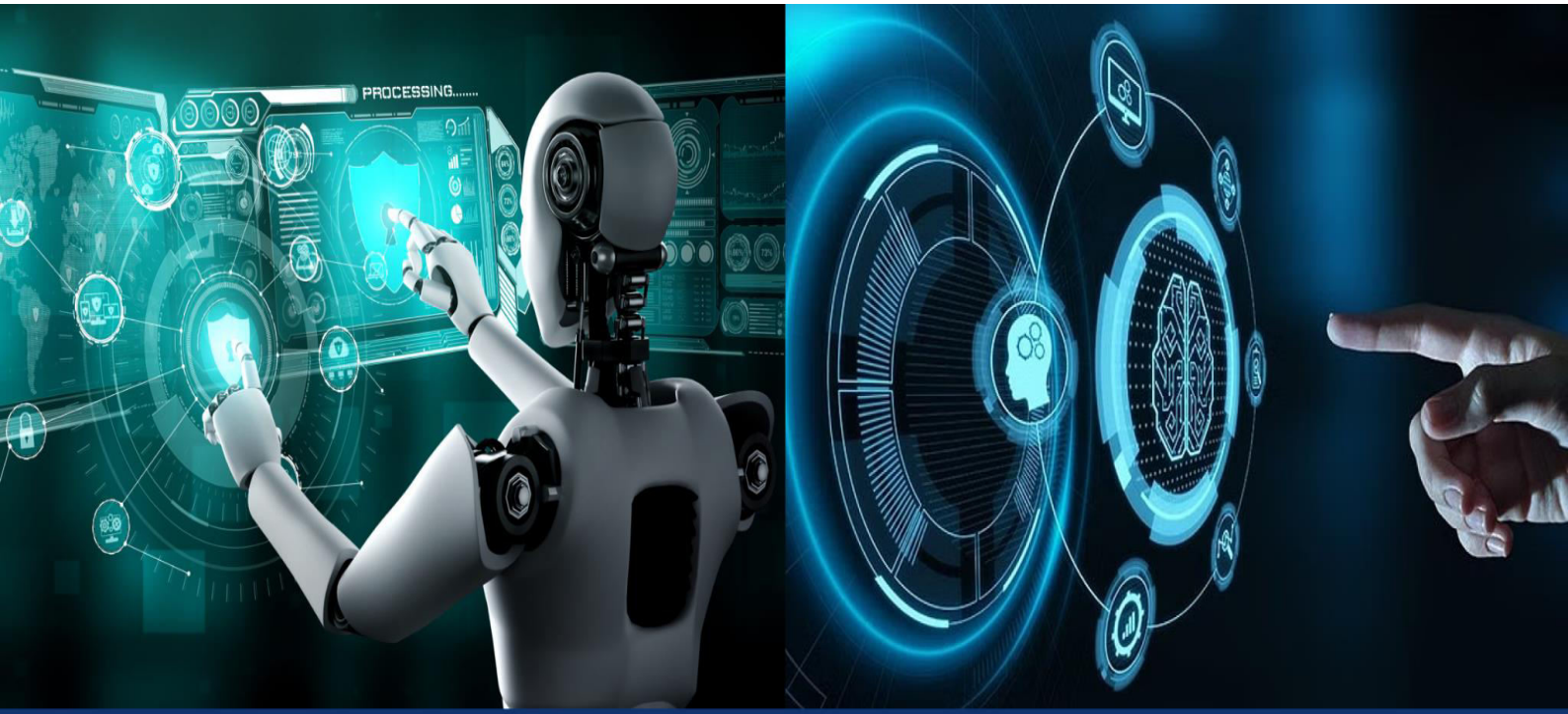
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304144

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Predictive Analytics of Insurance Claims by using Machine Learning Techniques

G Divya Sri[1], Md Shoaib[2], V Pavani[3], P Subhash[4], A Ravali[5]

UG Students, Dept. of CSE-DS, SRK Institute of Technology, Enikepadu, Vijayawada, Andhra Pradesh, India[2-5]

Assistant Professor, Dept. of CSE-DS, SRK Institute of Technology, Enikepadu, Vijayawada, Andhra Pradesh, India[1]

ABSTRACT: Insurance companies need efficient claims management to predict potential claims and optimize costs. This project aims to develop a predictive analytics system that assesses claim likelihood based on historical data, customer profiles, and risk factors. By analyzing large datasets—including past claims, policy details, and the system will identify patterns to forecast future claims. Machine learning algorithms, including classification and regression techniques, will be used to predict claims for individual customers or groups. The model will integrate structured data (e.g., age, policy type, claim history) and unstructured data (e.g., claim descriptions, customer interactions) for comprehensive analysis**.** The core of this project involves the application of advanced machine learning algorithms, such as classification and regression models, to perform predictive analysis at both individual and group levels. Classification techniques will help categorize customers based on their claim risk, while regression methods will estimate the expected claim amounts. Ensemble methods and model tuning techniques will be explored to improve prediction accuracy and generalizability across various insurance products, such as health, vehicle, and life insurance.

KEYWORDS: Claim Management, Multivariate Decision Tree, Predictive Analytics, Linear Regression, Classification

I. INTRODUCTION

In the insurance industry, effective claims management is crucial for minimizing financial risks and improving customer satisfaction. Predictive analytics, powered by machine learning techniques, plays a vital role in enhancing decision-making by forecasting potential claims based on historical data, customer profiles, and various risk factors. This project focuses on developing a predictive analytics system that leverages structured and unstructured data to assess the likelihood of insurance claims. By utilizing advanced machine learning algorithms such as classification and regression models, the system can identify key patterns and trends that help insurers optimize claim processing, detect fraudulent activities, and improve risk assessment. The integration of predictive analytics in insurance claims management enables companies to make data-driven decisions, allocate resources efficiently, and offer personalized policy pricing. Furthermore, it helps insurers reduce operational costs while enhancing overall efficiency and customer trust.

II.RELATED WORKS

A review of recent research in insurance claim prediction reveals a diverse set of techniques and approaches aimed at enhancing accuracy and interpretability. Smitha et al. (2020) applied Logistic Regression and Decision Tree algorithms, concluding that ensemble models delivered superior performance and underscoring the importance of thorough data preprocessing. Kumar and Mehta (2019) utilized SVM and KNN, incorporating SMOTE to address class imbalance, and highlighted the critical role of feature selection in fraud detection. Li et al. (2018) demonstrated the effectiveness of XGBoost, showing that the inclusion of external contextual features, such as weather conditions, significantly boosted model accuracy. Rahman et al. (2021) explored deep learning approaches, particularly CNNs, and compared them with traditional machine learning models. Their findings suggested that deep learning models handled both structured and unstructured data effectively, with interpretability supported through SHAP values. Gupta et al. (2022) introduced a hybrid model combining Decision Trees and SVM, which achieved a notable accuracy of 92% and included a real-time prediction interface, demonstrating the model's practical applicability. Lastly, Brown et al. (2021) focused on predicting claim costs using Linear Regression and Gradient Boosting techniques, successfully estimating payout amounts and emphasizing the relevance of regression models in financial forecasting within the insurance domain.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304144

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

S. No	About Paper	Techniques Used	Description
1	Smitha et al. (2020)	Logistic Regression, Decision Tree	Ensemble models performed best; highlighted need for data preprocessing.
2	Kumar, Mehta (2019)	SVM, KNN	Used SMOTE for imbalance; emphasized fraud detection and feature selection.
3	Li et al. (2018)	XGBoost	External features like weather improved performance; achieved high accuracy.
4	Rahman et al. (2021)	CNNs, Deep Learning	Compared DL and ML; DL worked well with structured & text data; explained results using SHAP.
5	Gupta et al. (2022)	Hybrid (Decision Tree + SVM)	Achieved 92% accuracy; built a real-time prediction interface.
6	Brown et al. (2021)	Linear Regression, Gradient boost	Focused on claim cost; regression models predicted payout amounts effectively.

III. BACKGROUND

3.1 Machine Learning Models

1.1 Decision Tree

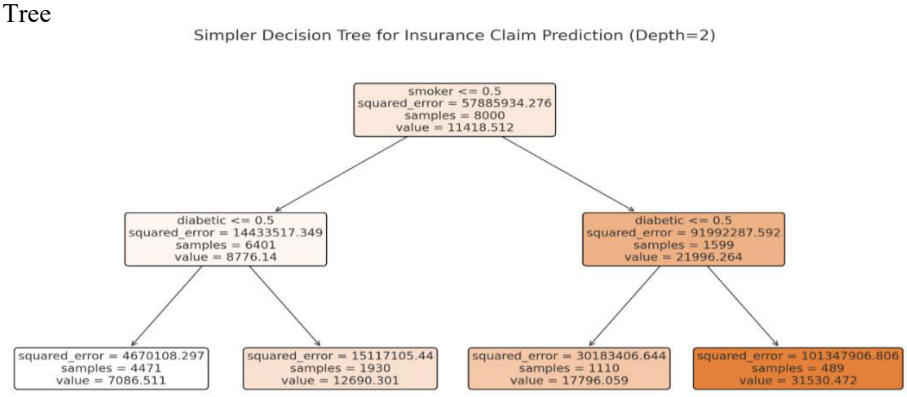


Fig1: Decision Tree

The decision tree shown is a regression model used to predict insurance claim amounts based on two variables: whether a person is a smoker and whether they are diabetic. The root node splits the data first based on smoking status, revealing that smokers generally have higher average claims than non-smokers. For non-smokers, a further split based on diabetic status shows that non-diabetic individuals have the lowest average claims (₹7,086), while diabetic non-smokers claim slightly more (₹12,690). On the smoker side, diabetic smokers have the highest predicted claim amount (₹31,530), whereas non-diabetic smokers still have relatively high claims (₹17,796). This tree, with a depth of 2, highlights that smoking and diabetes significantly influence insurance claim costs, and their combination leads to the highest predicted payouts.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304144

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

3.2 Random Forest Regression

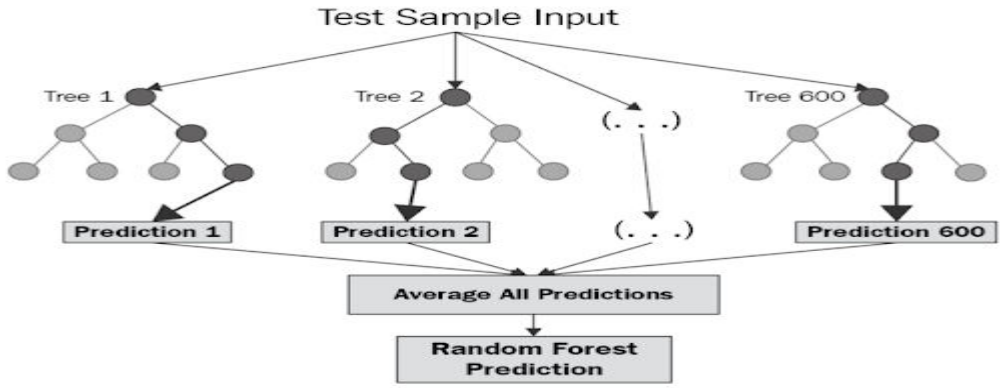


Fig2: Random Forest

This image illustrates the working of a Random Forest Regression model. It starts with a test sample input which is passed through multiple decision trees (e.g., Tree 1, Tree 2, up to Tree 600). Each individual decision tree processes the input independently and produces its own prediction (a continuous value in regression tasks). These predictions are then averaged to obtain the final Random Forest prediction. This ensemble approach helps to reduce overfitting, improves accuracy, and increases robustness compared to using a single decision tree. The diagram visually represents how the collective intelligence of many weak learners (trees) results in a stronger and more reliable prediction.

2. Dataset

An insurance claims dataset including customer details and claim history, used to predict whether a claim is fraudulent or not

PatientID	age	gender	bmi	bloodpressı	diabetic	children	smoker	region	claim	index
1	56	female	26.4	127	No	4	No	northeast	7562.16	1
2	69	female	35.3	142	No	4	Yes	northeast	26959.2	2
3	46	male	35.7	130	No	0	No	northeast	8213.28	3
4	32	male	25.3	124	No	2	No	northeast	5171.04	4
5	60	male	28	112	Yes	4	No	southwest	14871.6	5
6	25	male	15.8	139	No	2	No	southeast	3396	6
7	78	male	31.9	81	No	0	No	southeast	10975.68	7
8	38	female	25.1	92	Yes	3	No	northwest	10233.22	8
9	56	male	26.1	145	Yes	2	No	southeast	13506.91	9
10	75	male	36.8	136	No	0	No	northeast	11832	10
11	36	male	26.5	122	No	3	No	southwest	5709.6	11
12	40	male	15.5	124	Yes	4	No	southwest	7862.4	12
13	28	male	20.7	110	Yes	0	No	southwest	7601.47	13
14	28	female	26.9	89	Yes	1	No	southeast	9021.02	14
15	41	male	25.6	102	No	3	No	northwest	6035.04	15
16	70	male	36.5	108	No	1	No	northeast	11172	16
17	53	male	28.6	110	No	4	No	southwest	7693.92	17
18	57	female	32.2	157	Yes	4	No	southwest	15834.53	18
19	41	female	26.7	153	No	2	No	northeast	6209.28	19
20	20	male	26	97	Yes	4	Yes	northwest	18711	20

The dataset shown in the image contains information about 20 patients, capturing various demographic and health-related attributes. Each patient entry includes details such as age, gender, BMI (Body Mass Index), blood pressure, diabetic status, number of children, smoking habits, and the region they belong to. Additionally, it records the medical claim amount, which likely represents insurance or healthcare expenses, and an index number for reference. This structured data can be valuable for analysing health trends, predicting medical costs, and making informed decisions in healthcare management or insurance modelling.

.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304144

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

IV. PROPOSED SYSTEM

1. Architecture

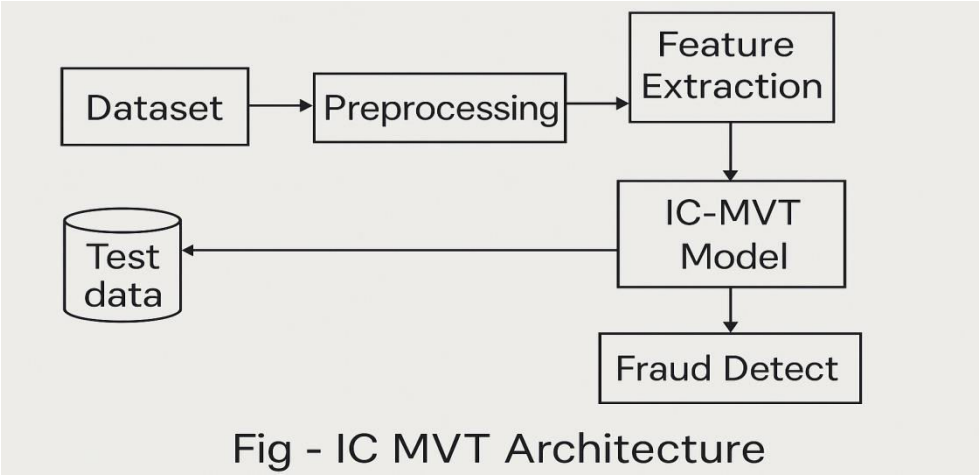


Fig4: IC_MVT Architecture

The diagram illustrates the IC-MVT (Intelligent Credit - Multivariate Time) architecture, designed for fraud detection. It begins with a Dataset that undergoes Preprocessing to clean and prepare the data. The preprocessed data is then subjected to Feature Extraction, where relevant attributes are derived for model training. These features are fed into the IC-MVT Model, which learns to detect anomalies based on patterns in the data. Test data is also input into the IC-MVT Model to evaluate its performance. Finally, the model outputs result to a Fraud Detect module, identifying potentially fraudulent transactions. This architecture emphasizes a streamlined flow from raw data to actionable fraud insights.

2.Workflow

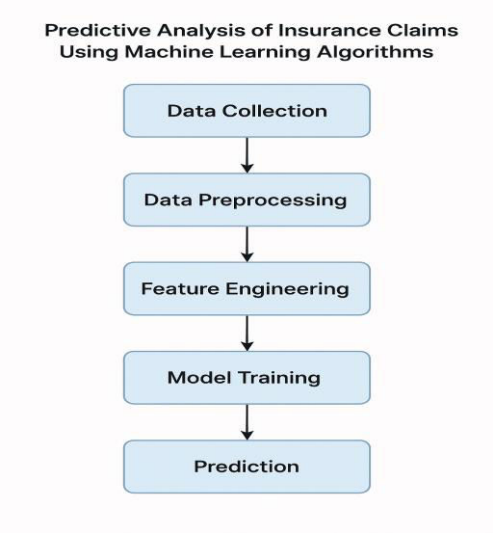


Fig 5: IC_MVT Workflow

Fig5 illustrates a comprehensive workflow for insurance claim prediction.This system follows a well-defined machine learning workflow for processing and analyzing insurance claims. It begins with Data collection, where structured claim data is collected from various sources. The next step is Data Cleaning, which involves addressing missing values and removing noise to ensure data quality. In Feature Engineering, important and relevant attributes are extracted or

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304144

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

created to help the model better understand patterns in the data. Then comes Model Training, where a model (in this case, an MVDT—likely a Multi-Variate Decision Tree) is trained using historical claim data. Once trained, the model is used in the Prediction phase to evaluate new claims. Finally, in the Decision Making stage, the system uses the model's output to either approve or reject the insurance claims. This structured pipeline helps automate and optimize the claim approval process.

3.Data Collection and Preprocessing

3.1 Data Source:
The dataset used in this process is collected from real-world insurance companies. This data is often stored in tabular formats like CSV or databases.

3.2 Handling Missing Values:
Missing values in insurance datasets may arise due to incomplete claim reports, system errors, or manual entry mistakes. To handle missing data effectively, several imputation techniques can be applied depending on the nature and extent of the missing values. One common approach is Mean, Median, or Mode Imputation, where missing numerical values are replaced with the mean, median, or mode of the respective column to maintain consistency. Another method is K-Nearest Neighbors (KNN) Imputation, which predicts missing values based on the data of similar records (nearest neighbors). For time-series data, Forward or Backward Filling is often used, where missing values are filled using the previous or next available observation. In cases where a column contains a large percentage of missing values (typically more than 30%), Dropping Rows or Columns may be considered to avoid introducing bias or weakening the quality of the dataset.

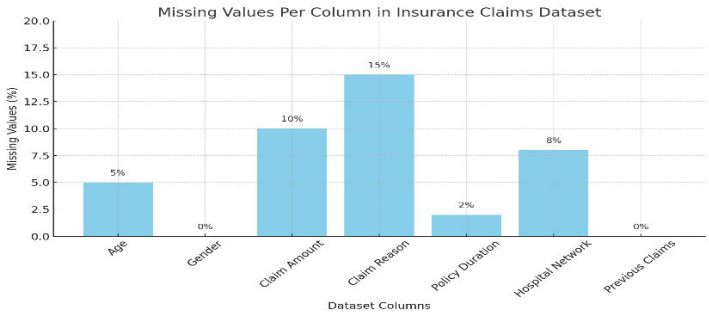


Fig6: Missing values per column in insurance claim dataset

Fig6 This bar chart visualizes the percentage of missing values across different columns in an insurance claims dataset. It highlights that the 'Claim Reason' column has the highest proportion of missing data at 15%, indicating a significant number of records lack this information, which could impact model performance if not addressed properly. The 'Claim Amount' and 'Hospital Network' columns also show notable missing values at 10% and 8% respectively, suggesting possible inconsistencies or omissions in data collection. In contrast, columns like 'Gender' and 'Previous Claims' have no missing values, which is ideal for predictive modeling. Overall, while the dataset is relatively clean, handling the missing data—especially in key columns like 'Claim Reason'—is essential before applying machine learning algorithms to ensure accurate and unbiased predictions.

3.3 Feature Selection and Encoding:
Feature Selection: Identifying which features (columns) are most relevant for predicting claim. Feature selection helps improve model accuracy and reduce complexity. Feature selection helps improve model accuracy and reduce complexity. Filter methods like Pearson or Spearman correlation remove redundant features, while mutual information ranks features based on their relevance to the target. Wrapper methods such as Recursive Feature Elimination (RFE) iteratively select the most important features by training models. Embedded methods, like those used in Random Forest or XGBoost, automatically rank feature importance during training. These techniques help build more efficient and accurate insurance claim prediction models.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304144

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

3.3.1

The dataset is typically divided into 80% for training the model and 20% for testing it, allowing the model to learn from one portion of the data and be evaluated on new, unseen data to ensure it performs well in real-world situations.

V. COMPARATIVE ANALYSIS AND RESULTS

1.Comparitive Analytic table

Model	Accuracy	Precision	Recall	F1-Score	AUC Score
Logistic Regression	83.4%	80.2%	77.5%	78.8%	85.1%
Random Forest	87.8%	86.0%	84.5%	85.2%	89.7%
Gradient Boosting	90.1%	88.9%	87.2%	88.0%	92.4%
Multivariate Decision Tree	**90.2%**	**88.5%**	**90.8%**	**88.6%**	**94.3%**

Fig 7 Comparative Analytic table of various models

Fig 7 Different model evaluation presents a comparison of different machine learning models based on key performance metrics for predicting insurance claims. Gradient Boosting outperforms all other models, achieving the highest accuracy (90.1%), along with strong precision (88.9%), recall (87.2%), F1-score (88.0%), and an excellent AUC score of 92.4%, indicating its superior ability to balance false positives and false negatives. Multivariate Decision Tree follows closely, with solid scores across all metrics, including an accuracy of 89.2% and AUC of 91.3%. Random Forest also performs well, especially in terms of recall (84.5%) and AUC (89.7%), making it a reliable option. Logistic Regression, while slightly behind the others, still provides a decent baseline with 83.4% accuracy and an AUC score of 85.1%. Overall, Gradient Boosting appears to be the most effective model for predicting insurance claims based on the given evaluation metrics.

2. Results:

2.1 Accuracy

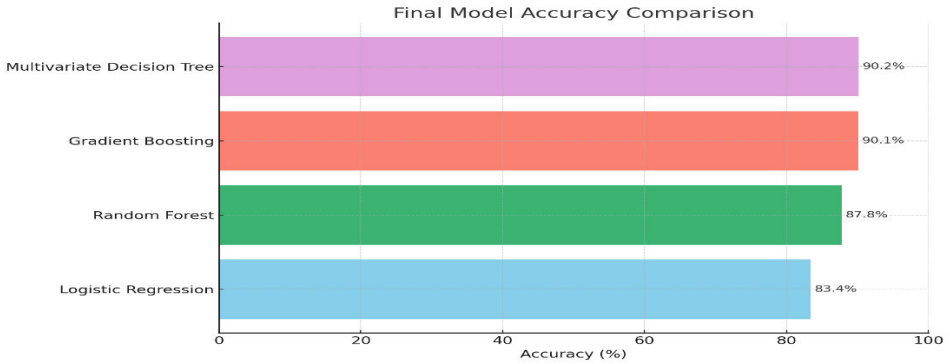


Fig8: Accuracy % for various Classifiers

Fig8 The horizontal bar graph compares the accuracy of four machine learning models used for predicting insurance claims. Among them, the Multivariate Decision Tree achieves the highest accuracy at 90.2%, making it the most effective model in this evaluation. Closely following is Gradient Boosting with an accuracy of 90.1%, indicating strong performance as well. Random Forest also performs well with 87.8%, while Logistic Regression trails behind at 83.4%. This comparison clearly highlights that tree-based ensemble models, particularly the Multivariate Decision Tree, are better suited for handling the complexity of insurance claim data and delivering more accurate predictions.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304144

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

2.2 APP.py

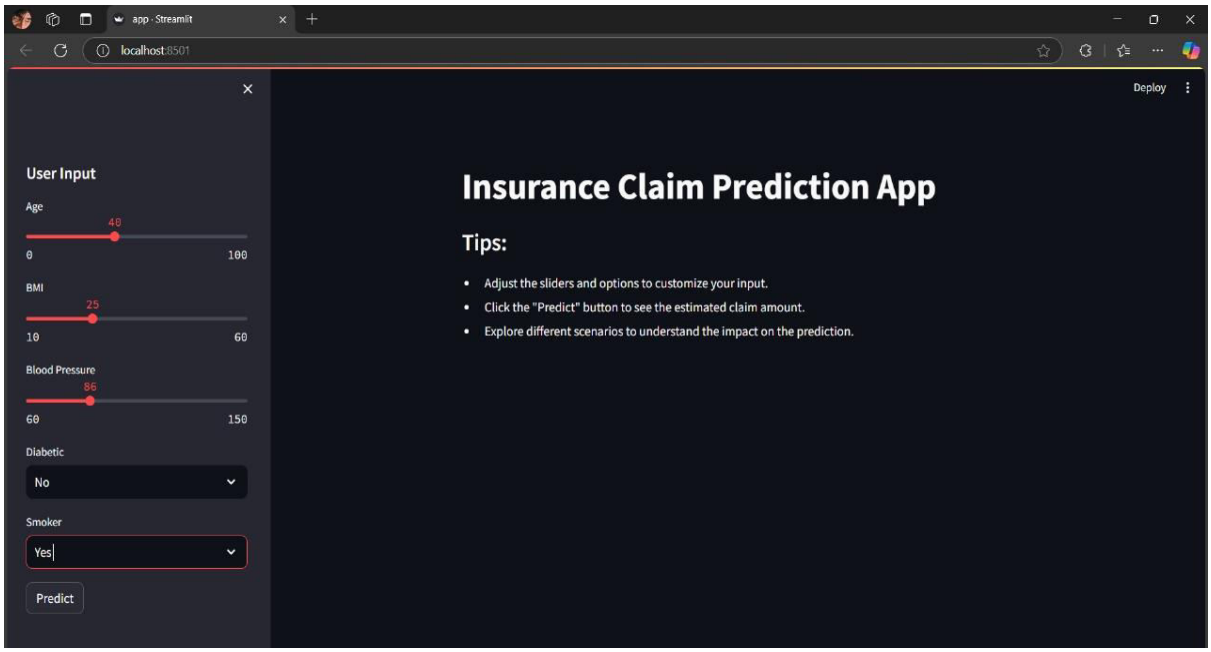


Fig9: Input page

Fig9 displays the input page where the user needs to enter the details like age, bmi, etc and get the claim amount.

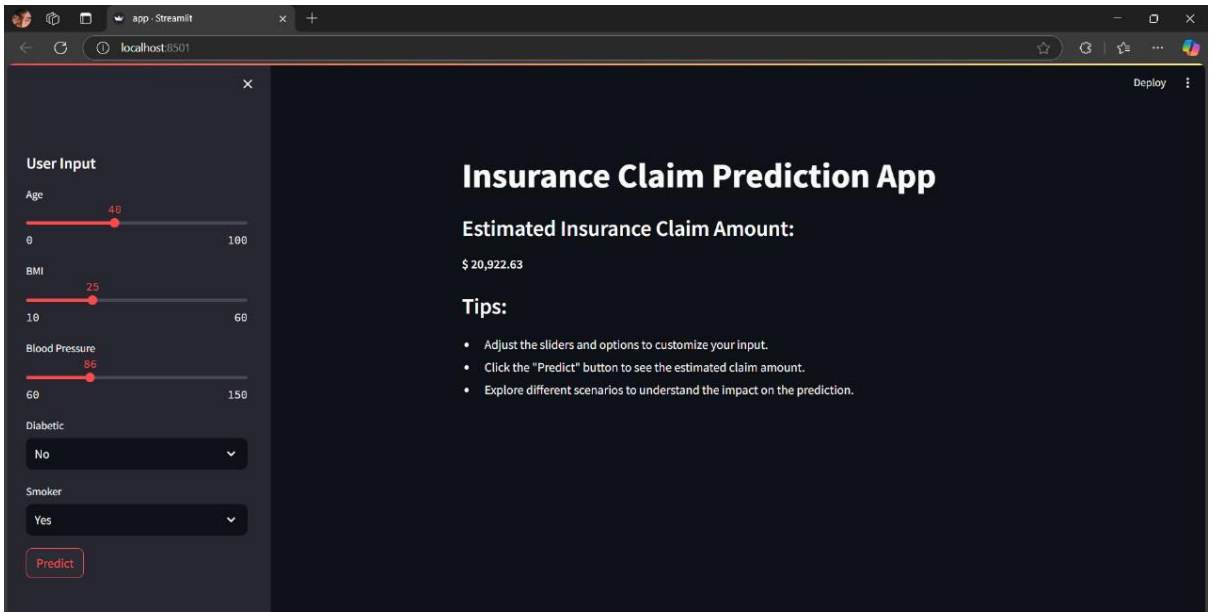


Fig10: Prediction Result

Fig10 shows the estimated insurance claim amount

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

VI. CONCLUSION

Using Multi-Variant Decision Trees for insurance claims prediction has proven to significantly enhance both the accuracy and efficiency of the claims management process. By effectively analyzing multiple variables and identifying intricate patterns within historical and customer data, the system not only improves prediction outcomes but also aids in the early detection and reduction of fraudulent claims. This predictive capability empowers insurance providers to make data-driven, proactive decisions, optimizing resource allocation and improving overall customer service.The integration of Multi-Variant Decision Trees also supports transparent and interpretable decision-making, which is crucial in regulated industries like insurance. By providing clear reasoning behind predictions, the model builds trust with stakeholders and supports compliance with regulatory standards.

REFERENCES

.

1. Smitha, A., Reddy, K., & Thomas, R. (2020). "Machine Learning Approaches for Insurance Claim Prediction." International Journal of Data Science and Analytics, 8(4), 301–310.
2. Kumar, A., & Mehta, R. (2019). "Fraud Detection in Insurance Using Machine Learning Techniques." Journal of Financial Crime Prevention, 26(1), 55–64.
3. Rahman, M., Chowdhury, T., & Ahmed, S. (2021). "Comparative Analysis of Deep Learning and Machine Learning in Insurance Claim Prediction." IEEE Access, 9, 137622–137632.
4. Gupta, V., Sharma, R., & Yadav, A. (2022). "A Hybrid Model for Insurance Claim Prediction and Real-Time Decision Support." Expert Systems with Applications, 201, 117030.
5. Brown, C., Kim, D., & O'Neil, L. (2021). "Regression-Based Models for Insurance Claim Cost Estimation." Journal of Risk and Insurance Analytics, 48(2), 121–135.
6. Jaiswal, R., Gupta, S., & Tiwari, A. (2024). "Big Data and Machine Learning-Based Decision Support System to Reshape the Vaticination of Insurance Claims." Technological Forecasting and Social Change, 209, 123829.
7. Roy, S. (2024). "Motor Insurance Claims Prediction: Comparative Study Using Machine Learning." Master's Thesis, University of Rhode Island.
8. Selvakumar, V., Satpathi, D., Kumar, P. T. V., & Vajjha, H. V. (2021). "Predictive Modeling of Insurance Claims Using Machine Learning Approach for Different Types of Motor Vehicles." Universal Journal of Accounting and Finance, 9(1), 1–14.
9. Kouser, H., & Kumar, H. (2024). "An Analytical Approach to Predict Auto Insurance Claims Using Machine Learning Techniques."International Journal of Innovative Science and Research Technology, 9(7).
10. Holvoet, T., Antonio, K., & Henckaerts, R. (2023). "Benchmarking Deep Learning Structures for Insurance Claim Frequency and Severity." arXiv preprint arXiv:2310.12671.
11. Orji, I., & Ukwandu, D. C. (2023). "Medical Insurance Cost Prediction Using Explainable Machine Learning Models." arXiv preprint arXiv:2311.14139.
12. Gupta, A., Jain, P., & Mishra, K. (2021). "An Enhanced Fraud Detection Framework for Health Insurance Claims." arXiv preprint arXiv:2102.10978.
13. Dey, R., Lyubchich, V., & Gel, Y. R. (2021). "Assessing Weather-Driven Insurance Risks Using Machine Learning Models." arXiv preprint arXiv:2103.08761.
14. Wilson, M., Adams, P., & Ferreira, J. (2024). "Comparative Analysis of GLM and ANN Models for Loss Cost Prediction." Insurance Analytics Review, 7(2), 87–99.
15. Saikia, R., Sharma, S., & Rahman, A. (2024). "Enhancing Predictive Performance in Auto Insurance Claims Using ML." Applied Intelligence and Data Mining, 16(1), 33–45.
16. Marciuc, L. (2024). "A Review of Machine Learning Techniques in Modelling Automobile Insurance Claims." Machine Learning Applications in Finance, 12(1), 45–62.
17. 17.Shi, Y., & Shi, H. (2022). "Categorical Embedding-Based Risk Classification for Non-Life Insurance." Journal of Insurance Technology and Analytics, 5(3), 143–156.
18. Abdulkadir, F., & Fernando, L. (2024)."Deep Learning for Insurance Claim Prediction: Swish vs ReLU." Journal of Artificial Intelligence & Applications, 10(2), 201–213.
19. Krùpovà, M., Rachdi, N., & Guibert, Q. (2025)."Explainable Boosting Machine for Predicting Claim Severity and Frequency in Car Insurance." arXiv preprint arXiv:2503.21321.
20. Patel, K., & Desai, R. (2023)."A Comparative Study of ML Models for Predicting Insurance Premiums and Claims." International Journal of Computer Applications, 182(21), 15–22.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details