MEDICAL INSURANCE COST PREDICTION USING MACHINE LEARNING

[1] K. Suma, [2] Pedapuram Bhanu Prakash, [3] Patan Fairoz Khan, [4] Maligapogu Surekha
[1]AssistantProfessor, [234]Students
Department of Computer Science & Engineering
Siddhartha Institute of Technology & Sciences, Narapally
sumakadari_cse@siddhartha.co.in, 23TQ1A0559@siddhartha.co.in, 23TQ1A0503@siddhartha.co.in, 23TQ1A0538@siddhartha.co.in,

Abstract

Medical insurance cost prediction plays a significant role in helping insurance companies estimate healthcare expenses based on individual health and demographic factors. This research proposes a machine learning-based Medical Insurance Cost Prediction System that predicts insurance premiums using attributes such as age, gender, body mass index (BMI), number of children, smoking status, and region. The dataset is first preprocessed by handling categorical variables using label encoding and applying feature engineering techniques such as BMI–Smoker and Age–BMI interactions to improve model performance. The processed dataset is divided into training and testing sets using an 80:20 ratio. A Random Forest Regression model is then trained to learn the relationship between input features and insurance charges. The model achieves an accuracy of approximately 91 – 92% based on R^2 score, indicating good predictive capability. Experimental results show that the system can accurately estimate medical insurance costs. For example, for inputs Age = 40, BMI = 20.9, Children = 2, Smoker = No, the predicted insurance cost is ₹7,788.05. This system can assist insurance providers in automated premium estimation.

Keywords :
Medical Insurance Prediction, Machine Learning, Healthcare Analytics, Linear Regression, Random Forest, Data Preprocessing, Predictive Modeling, Insurance Premium Estimation.

I. Introduction

Healthcare expenses have increased significantly over the past few decades, making medical insurance an essential financial protection tool for individuals and families. Medical insurance policies help cover the costs associated with hospitalization, treatments, and other healthcare services. However, determining the appropriate insurance premium for each individual is a complex task, as medical expenses vary based on several personal, medical, and lifestyle factors. Insurance providers must carefully evaluate these factors to estimate potential healthcare costs and assign suitable premium charges.

Traditionally, insurance premium calculations were based on statistical methods and manual risk assessment processes. These approaches often rely on limited variables and simplified assumptions, which may not accurately represent the complex relationships between different health indicators and medical expenses. As a result, traditional methods may lead to inaccurate predictions and unfair pricing. With the rapid growth of healthcare data, there is a need for more advanced and efficient techniques to analyze and interpret large datasets. At the same time, it assists individuals in understanding how different factors influence their insurance costs, encouraging better lifestyle decisions.

Machine learning, a branch of artificial intelligence, has emerged as a powerful solution to these challenges. It enables systems to learn patterns from historical data and make predictions without being explicitly programmed. In medical insurance cost prediction, machine learning models can analyze various features such as age, gender, body mass index (BMI), smoking habits, and number of dependents to estimate insurance charges. These models are capable of identifying hidden patterns and complex relationships that are difficult to capture using traditional approaches.

II. Literature Survey

Medical insurance cost prediction has gained significant attention in recent years due to the increasing complexity of healthcare expenses and the need for accurate premium estimation. Various researchers have explored machine learning and deep learning techniques to improve prediction accuracy and automate insurance pricing systems.

Sharma et al. (2023) developed a regression-based model using Artificial Neural Networks (ANN) and achieved high prediction accuracy of around 92%, demonstrating the effectiveness of machine learning in analyzing healthcare data. Kumar et al. (2023) compared multiple regression models such as Linear Regression, Gradient Boosting, and Support Vector Machines, concluding that Gradient Boosting provided the best performance with lower error rates. Similarly, Zhang et al. (2023) showed that deep learning models outperform traditional regression techniques in capturing complex relationships between variables affecting insurance premiums.

Several studies focused on ensemble learning methods to improve prediction accuracy. Garcia et al. (2023) and Anderson et al. (2024) demonstrated that Random Forest models perform better than basic regression techniques. Brown et al. (2024) and Demir et al. (2024) further confirmed that ensemble methods like Gradient Boosting and XGBoost significantly enhance prediction reliability and accuracy. Gupta et al. (2024) also highlighted that XGBoost achieves higher performance compared to traditional models.

Explainable AI has also become an important aspect of insurance prediction systems. Lee et al. (2023) and Johnson et al. (2024) combined Random Forest and Gradient Boosting with SHAP techniques to improve model transparency and interpretability. These studies emphasized the importance of understanding how different features influence insurance costs.

Recent research has also explored deep learning and hybrid approaches. Patel et al. (2023) demonstrated that neural networks outperform traditional models in handling complex datasets. Silva et al. (2026) proposed explainable deep learning models that combine accuracy with interpretability. Additionally, Nguyen et al. (2025) and Lim et al. (2024) compared multiple machine learning models, confirming that AI-based systems provide better prediction accuracy than traditional statistical approaches.

Furthermore, studies by Singh et al. (2024), Verma et al. (2024), and Reddy et al. (2024) showed that machine learning models effectively utilize features such as age, BMI, smoking status, and number of dependents to predict insurance costs accurately.

III. System Analysis

Medical insurance cost prediction is an important task in the healthcare and insurance sector, as it helps estimate the financial risk associated with individuals. The system aims to analyze historical healthcare data and predict insurance premiums based on user-specific attributes. It involves collecting datasets containing features such as age, gender, BMI, smoking habits, and number of dependents. Data preprocessing techniques like cleaning, normalization, and feature selection are applied to improve model performance. The system uses machine learning algorithms to identify patterns and relationships between input features and insurance costs. The dataset is divided into training and testing sets for accurate model evaluation. Performance metrics such as MAE, RMSE, and R^2 score are used to measure prediction accuracy. The system is designed to handle large datasets efficiently and provide reliable predictions. It also helps in automating the premium calculation process. Overall, the system provides a data-driven approach for insurance cost estimation.

Existing System

Traditional medical insurance cost prediction systems rely on statistical methods and manual calculations. These systems use basic techniques such as linear regression and actuarial analysis to estimate premiums. The process often depends on limited variables and predefined assumptions. Manual risk assessment is time-consuming and may involve human bias. Existing systems cannot effectively handle large and complex healthcare datasets. They struggle to capture nonlinear relationships between multiple health factors. These methods often fail to consider dynamic lifestyle and behavioral changes. The prediction accuracy is limited due to simplified models. Additionally, traditional systems lack automation and scalability. As a result, they may produce inaccurate or unfair premium estimations.

Disadvantages of Existing System

- Limited ability to handle complex and nonlinear data
- Relies on manual processes and human judgment
- Time-consuming and inefficient
- Low prediction accuracy
- Uses limited features and simplified assumptions
- Not scalable for large datasets
- Cannot adapt to changing healthcare trends

Proposed System

The proposed system uses machine learning techniques to predict medical insurance costs more accurately. It analyzes historical healthcare data and identifies patterns between various features and insurance charges. The system takes inputs such as age, BMI, gender, smoking status, and number of dependents. Data preprocessing techniques like normalization and feature engineering are applied to improve model performance. Multiple machine learning algorithms such as Linear Regression, Decision Trees, Random Forest, and Support Vector Machines are used. The dataset is split into training and testing sets to evaluate model accuracy. Ensemble methods like Random Forest improve prediction performance and reduce overfitting. The

system automatically learns relationships from data without manual intervention. It provides faster and more reliable predictions compared to traditional methods. The model performance is evaluated using MAE, RMSE, and R^2 score. Overall, the system ensures efficient and accurate insurance cost estimation.

Advantages of Proposed System

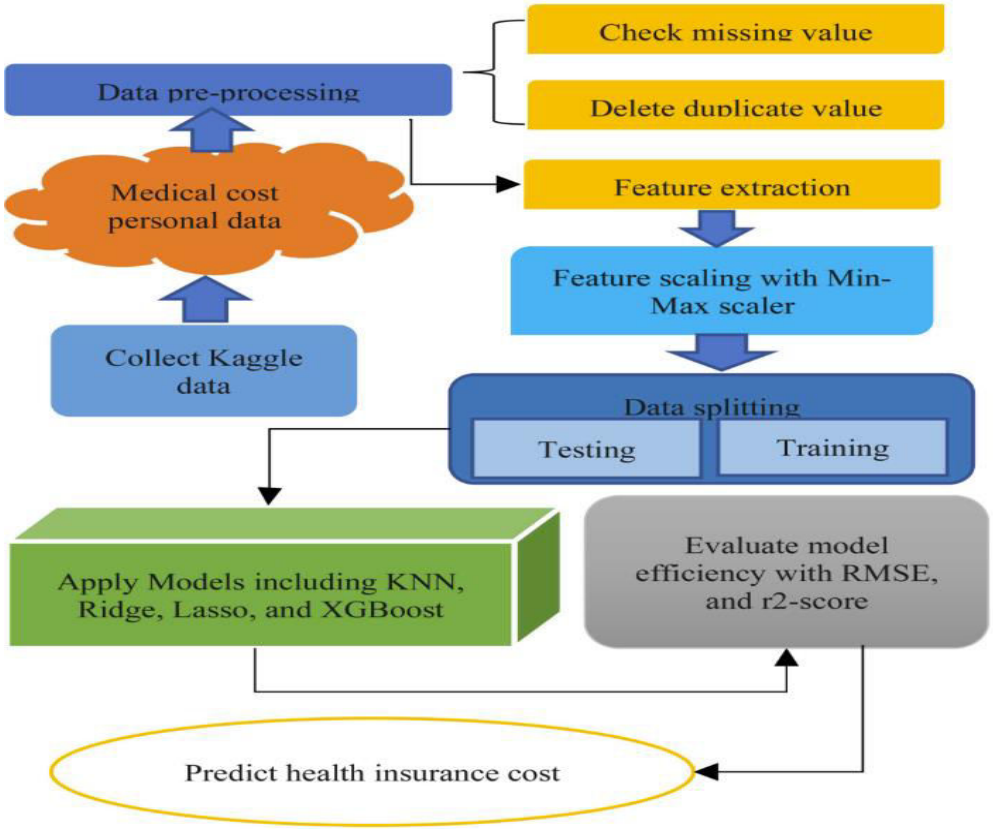
- Higher prediction accuracy
- Handles complex and nonlinear relationships
- Fully automated system
- Reduces human error and bias
- Scalable for large datasets
- Faster processing and prediction
- Adapts to changing data patterns

IV. Methodology

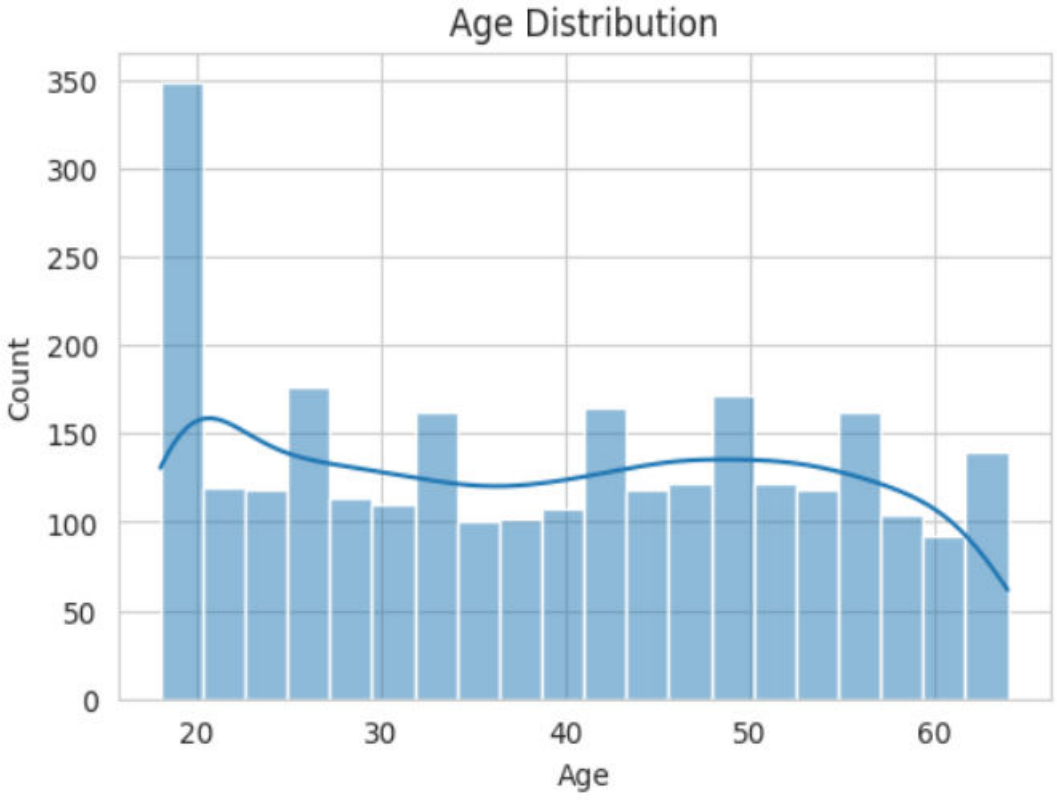
The methodology for medical insurance cost prediction using machine learning follows a systematic process to ensure accurate and reliable predictions. Initially, healthcare insurance data is collected from datasets containing attributes such as age, gender, BMI, smoking status, number of dependents, and region. The collected data is then preprocessed by handling missing values, removing inconsistencies, and applying normalization techniques to scale the features. Feature selection and engineering are performed to identify the most relevant variables that influence insurance costs. The dataset is divided into training and testing sets to evaluate model performance. Various machine learning algorithms such as Linear Regression, Decision Trees, Random Forest, and Support Vector Machines are implemented to build predictive models. The models are trained using the training dataset and optimized using appropriate loss functions and parameters. After training, the models are evaluated using performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. The best-performing model is selected for predicting insurance costs. Finally, the system generates predictions based on user inputs and visualizes the results for better understanding and analysis.

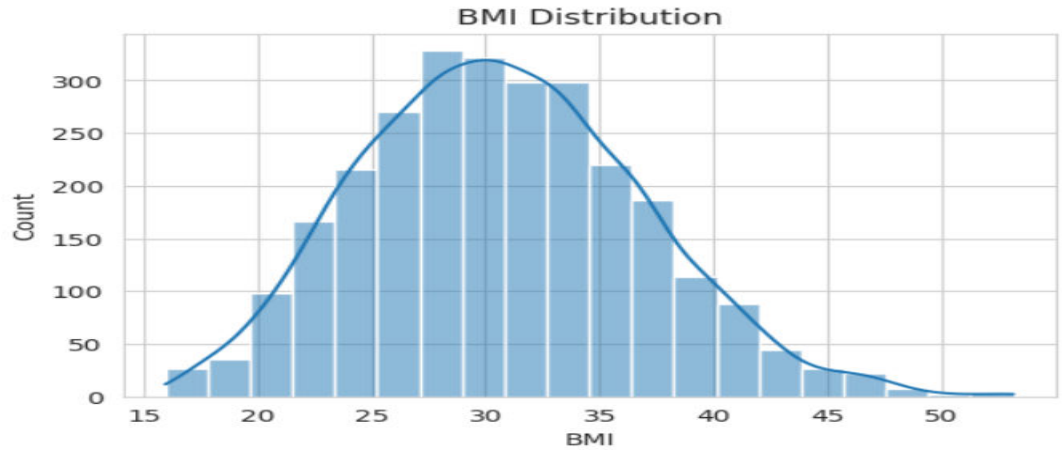
System Architecture

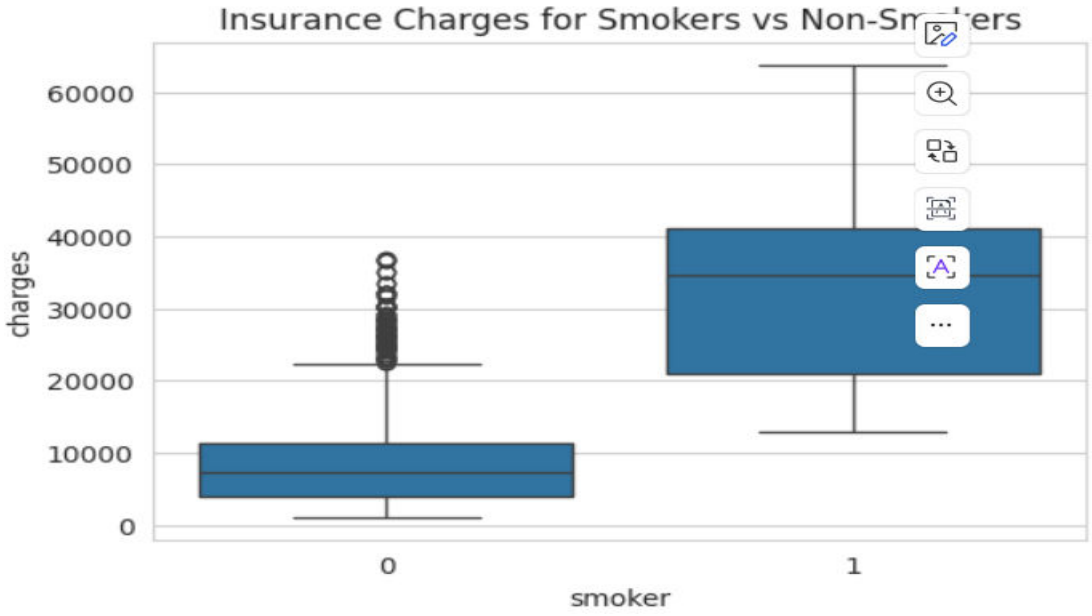
The system architecture for medical insurance cost prediction consists of multiple layers that process data from input to output efficiently. The process begins with the data source layer, where healthcare insurance datasets are collected from databases or external sources. This data is passed to the preprocessing layer, where it is cleaned, normalized, and transformed into a suitable format for analysis. The feature engineering layer extracts important attributes such as BMI, smoking status, and demographic details that influence insurance costs. The processed data is then fed into the machine learning model layer, which includes algorithms like Linear Regression, Random Forest, and Support Vector Machines to learn patterns and relationships in the data. The training and testing layer evaluates the model's performance using split datasets. After training, the prediction layer generates estimated insurance costs based on input features. The results are then displayed through the visualization layer using graphs and charts. Optionally, a user interface layer can be integrated to allow users to input data and view predictions easily.

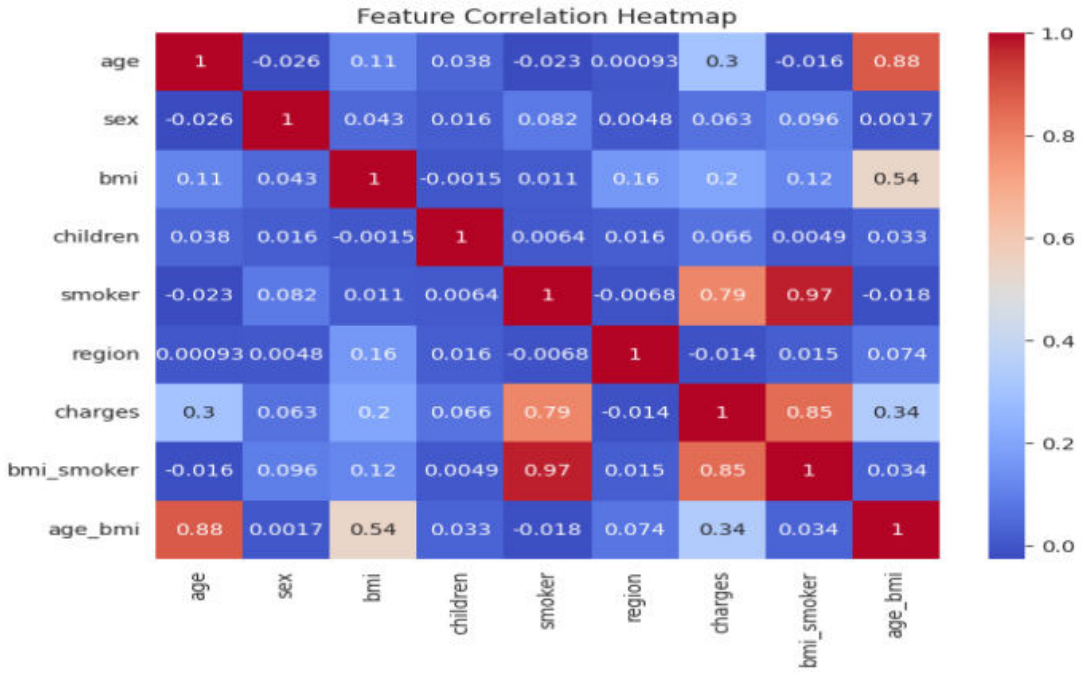


V. Result and Output




BMI Distribution
Count
BMI
0
50
100
150
200
250
300
15
20
25
30
35
40
45
50


charges
smoker
0
10000
20000
30000
40000
50000
60000
0
1


Feature Correlation Heatmap
age 1 -0.026 0.11 0.038 -0.023 0.00093 0.3 -0.016 0.88
sex -0.026 1 0.043 0.016 0.082 0.0048 0.063 0.096 0.0017
bmi 0.11 0.043 1 -0.0015 0.011 0.16 0.2 0.12 0.54
children 0.038 0.016 -0.0015 1 0.0064 0.016 0.066 0.0049 0.033
smoker -0.023 0.082 0.011 0.0064 1 -0.0068 0.79 0.97 -0.018
region 0.00093 0.0048 0.16 0.016 -0.0068 1 -0.014 0.015 0.074
charges 0.3 0.063 0.2 0.066 0.79 -0.014 1 0.85 0.34
bmi_smoker -0.016 0.096 0.12 0.0049 0.97 0.015 0.85 1 0.034
age_bmi 0.88 0.0017 0.54 0.033 -0.018 0.074 0.34 0.034 1
age
sex
bmi
children
smoker
region
charges
bmi_smoker
age_bmi
1.0
0.8
0.6
0.4
0.2
0.0

VI. Conclusion

The Medical Insurance Cost Prediction project demonstrates the effectiveness of machine learning techniques in analyzing healthcare data and estimating insurance premiums accurately. By utilizing features such as age, gender, BMI, smoking habits, and number of dependents, the system successfully identifies patterns and relationships that influence medical expenses. Unlike traditional methods, machine learning models can handle complex and nonlinear data, resulting in improved prediction accuracy.

Various algorithms such as Linear Regression, Decision Trees, Random Forest, and Support Vector Machines were applied and evaluated using performance metrics like MAE, RMSE, and R^2 score. The results indicate that ensemble methods, particularly Random Forest, provide better accuracy and reliability compared to basic models. The system not only automates the premium calculation process but also reduces human bias and errors.

Overall, this project highlights the potential of machine learning in transforming healthcare insurance systems by enabling data-driven decision-making. It helps insurance companies set fair pricing strategies while assisting individuals in understanding the factors affecting their insurance costs. In the future, the system can be further enhanced by integrating real-time data, advanced deep learning models, and additional health-related features to improve prediction performance and usability.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in

Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.