MEDICAL INSURANCE COST PREDICTION

[1] P. Yamini Chouhan, [2] G. Sahana, [3] K. Navadeep, [4] M. Nikhil
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
yaminichouhan_cse@siddhartha.co.in, 23tq1a5630@siddhartha.co.in, 23tq1a5629@siddhartha.co.in, 23tq1a5659@siddhartha.co.in

Abstract

Medical insurance costs are influenced by a wide range of demographic, lifestyle, and health-related factors, making accurate prediction a complex task. This project aims to develop a machine learning-based model to estimate individual medical expenses using features such as age, gender, body mass index (BMI), number of children, smoking status, and region. A publicly available dataset containing 2,772 records is utilized for analysis and model development.

The process begins with data preprocessing and exploratory data analysis to understand patterns and relationships within the dataset. A **Random Forest Regressor** is then implemented to model the relationship between input features and insurance charges. The model achieves a strong performance with an **R^2 score of 0.885** on the test dataset, indicating its effectiveness in explaining approximately 89% of the variance in medical costs.

Feature importance analysis reveals that smoking status, age, and BMI are the most significant factors influencing insurance charges. The results demonstrate the capability of machine learning techniques to provide accurate and reliable cost predictions. This system can assist insurance companies in risk assessment, premium calculation, and policy planning, while also helping individuals better understand and anticipate their medical expenses.

I. Introduction

In the modern healthcare industry, medical insurance plays a vital role in providing financial security against rising healthcare costs. However, determining accurate insurance premiums is a complex task due to the influence of multiple factors such as age, lifestyle habits, medical conditions, and geographic location. Traditional methods of estimating insurance costs are often based on generalized assumptions and static models, which may not capture the true risk associated with each individual.

With the advancement of technology, **Machine Learning (ML)** has emerged as a powerful tool for analyzing large datasets and identifying hidden patterns. By leveraging ML techniques, it is possible to build predictive models that can estimate medical insurance costs more accurately based on individual characteristics. These models consider various features such as age, gender, body mass index (BMI), number of dependents, smoking habits, and region to provide personalized predictions.

This project focuses on developing a machine learning-based system using regression techniques to predict medical insurance charges.

II. Literature Survey

Medical insurance cost prediction has gained significant attention in recent years due to the growing need for accurate and fair pricing in the healthcare industry. Early studies in this domain primarily relied on traditional statistical methods such as linear regression, which provided a simple approach to understanding the relationship between factors like age, BMI, and medical expenses. However, these methods often struggled to capture complex, non-linear relationships present in real-world data.

With the advancement of machine learning, researchers have explored various algorithms such as Decision Trees, Random Forest, Support Vector Regression (SVR), and Gradient Boosting to improve prediction accuracy. Among these, Random Forest has been widely adopted due to its ability to handle non-linearity, reduce overfitting, and provide feature importance insights. Studies have shown that ensemble methods like Random Forest and Gradient Boosting outperform traditional regression models in terms of accuracy and robustness.

Recent research has also investigated the use of deep learning techniques, including Artificial Neural Networks (ANNs), to model complex interactions between features. These approaches have demonstrated improved predictive performance, especially when dealing with large and diverse datasets. However, they often require more computational resources and careful tuning compared to traditional machine learning models.

A key finding across multiple studies is the strong influence of certain features—particularly smoking status, age, and BMI—on medical insurance costs. Researchers consistently highlight these variables as the most significant predictors of healthcare expenses. Additionally, issues such as data imbalance, feature correlation, and model interpretability are commonly discussed challenges in this field.

III. System Analysis

The existing system for medical insurance cost prediction relies on traditional statistical methods and fixed pricing models based on limited factors. These systems use simple approaches like linear regression and manual evaluation, which lack accuracy and fail to capture complex relationships. As a result, they often produce unfair or generalized insurance premiums. The disadvantages of the existing system include low prediction accuracy, inability to handle large datasets, lack of adaptability, and poor consideration of lifestyle factors such as smoking and BMI.

To overcome these limitations, the proposed system uses machine learning techniques, specifically a Random Forest Regressor, to predict insurance costs more accurately. It analyzes multiple features such as age, BMI, smoking status, and region to provide personalized predictions. The advantages of the proposed system include higher accuracy, better handling of non-linear data, scalability, and improved decision-making. System analysis involves understanding the current methods of predicting medical insurance costs, identifying their limitations, and proposing an improved system using machine learning techniques.

Existing System

The existing system for predicting medical insurance costs is primarily based on traditional statistical methods and manual evaluation techniques. Insurance companies generally use fixed pricing models, historical data analysis, and simple algorithms such as linear regression to estimate insurance premiums. These systems consider only a limited number of factors like age, basic medical history, and past claims, often ignoring important lifestyle and behavioral factors such as smoking habits, BMI, and region.

Additionally, the existing approach relies heavily on generalized assumptions rather than personalized data, which leads to less accurate predictions. These systems are not capable of capturing complex and non-linear relationships between different variables affecting insurance costs. As a result, they may produce biased or unfair premium estimates and are not well-suited for handling large and diverse datasets in modern healthcare environments.

Disadvantages of Existing System

- Lack of accuracy due to oversimplified models
- Inability to capture complex and non-linear relationships
- Limited consideration of important features like lifestyle factors
- High chances of unfair premium calculation
- Not adaptable to new or unseen data patterns
- Time-consuming manual analysis

Proposed System

The proposed system uses **Machine Learning techniques**, specifically a **Random Forest Regressor**, to accurately predict medical insurance costs. Unlike traditional methods, it considers multiple features such as age, gender, BMI, number of children, smoking status, and region to provide personalized predictions. The system performs data preprocessing, feature selection, and model training to learn patterns from historical data. It is capable of capturing complex and non-linear relationships between variables, leading to improved prediction accuracy. The model can also handle large datasets efficiently and adapt to new data over time, making it suitable for real-world applications.

Advantages of Proposal System

- High prediction accuracy compared to traditional methods
- Handles complex and non-linear relationships effectively
- Provides personalized insurance cost estimation
- Considers multiple important factors (age, BMI, smoking, etc.)
- Scalable for large datasets
- Reduces risk of unfair premium calculation
- Faster and automated decision-making

IV. Methodology

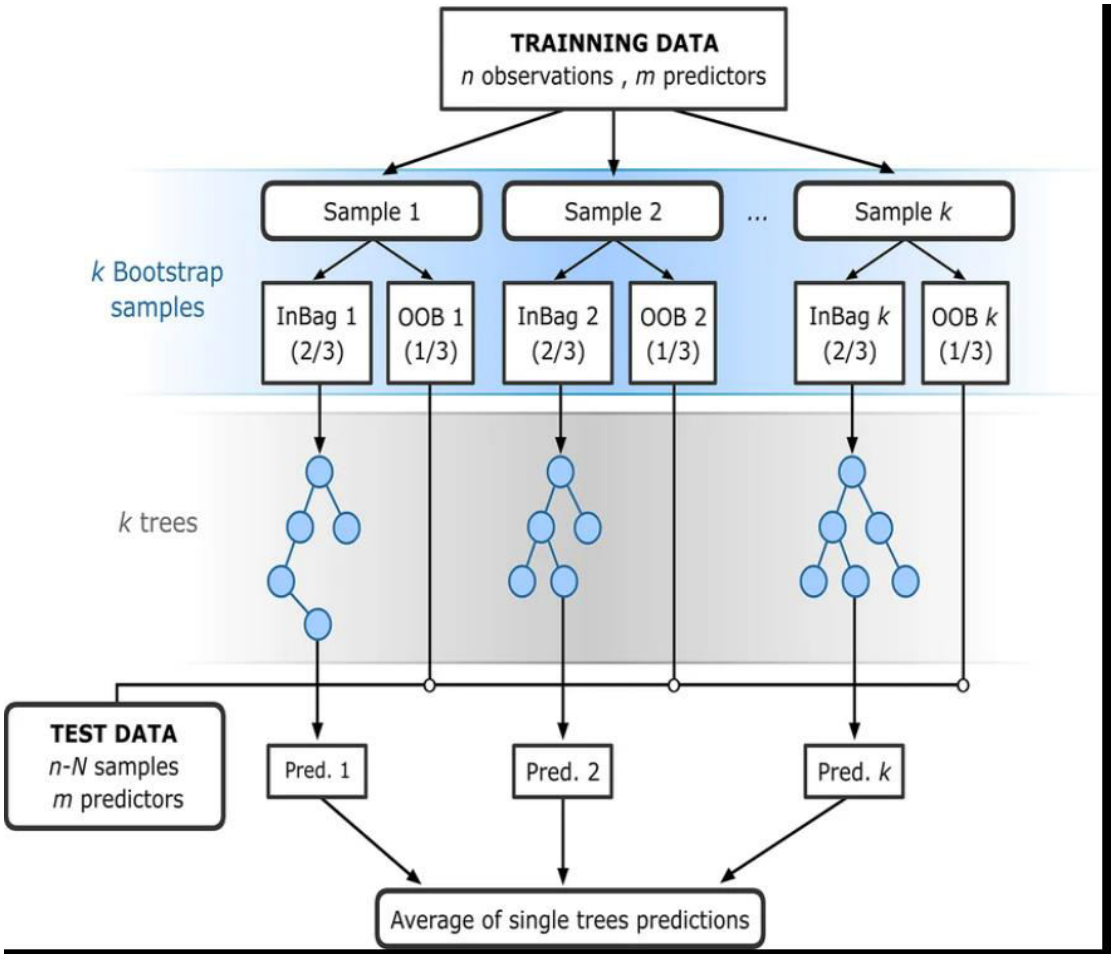
The methodology for predicting medical insurance costs involves several systematic steps to build an accurate and efficient machine learning model. Initially, a publicly available dataset containing medical and demographic details is collected. The dataset includes features such as age, gender, BMI, number of children, smoking status, and region, along with corresponding insurance charges.

The next step is data preprocessing, where missing values (if any) are handled, and categorical variables such as sex, smoking status, and region are converted into numerical form using encoding techniques. Data cleaning is also performed to remove inconsistencies and improve data quality.

Following preprocessing, exploratory data analysis (EDA) is conducted to understand relationships between variables and identify important features affecting insurance costs. Visualization techniques help in analyzing trends and correlations.

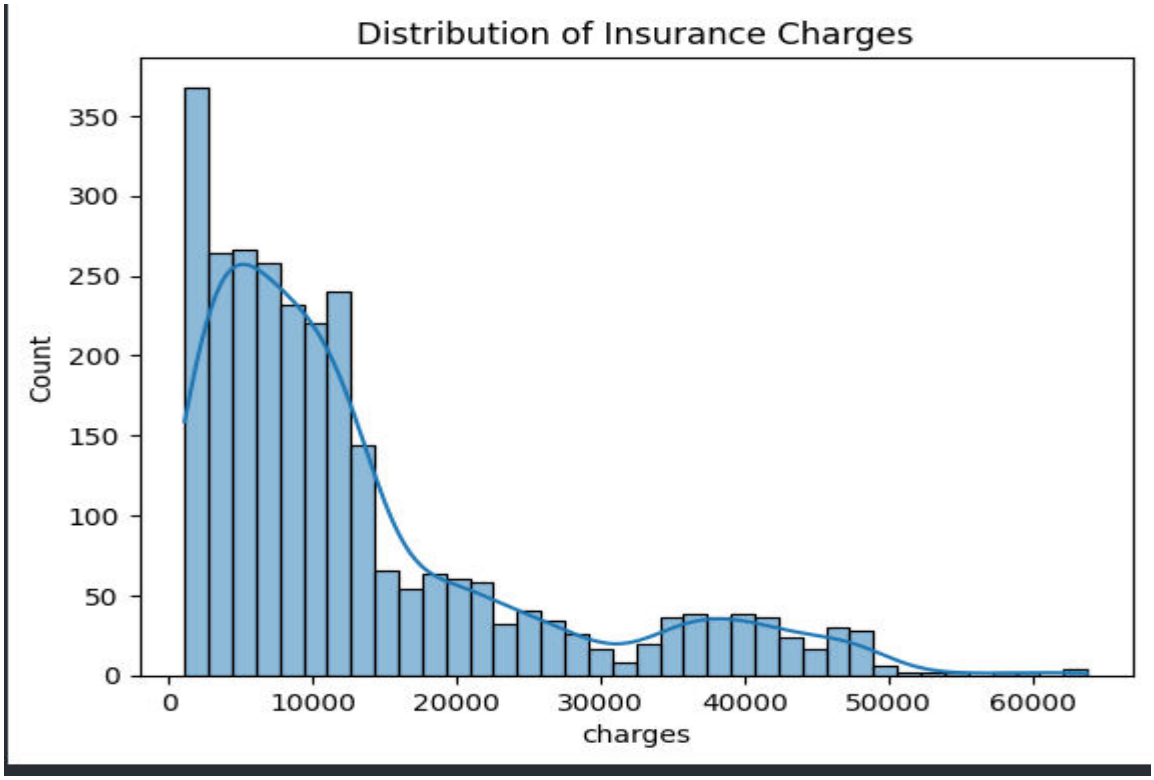
System Architecture

The system architecture defines how different components of the medical insurance cost prediction system interact to produce accurate results. It follows a structured pipeline from data input to final prediction.

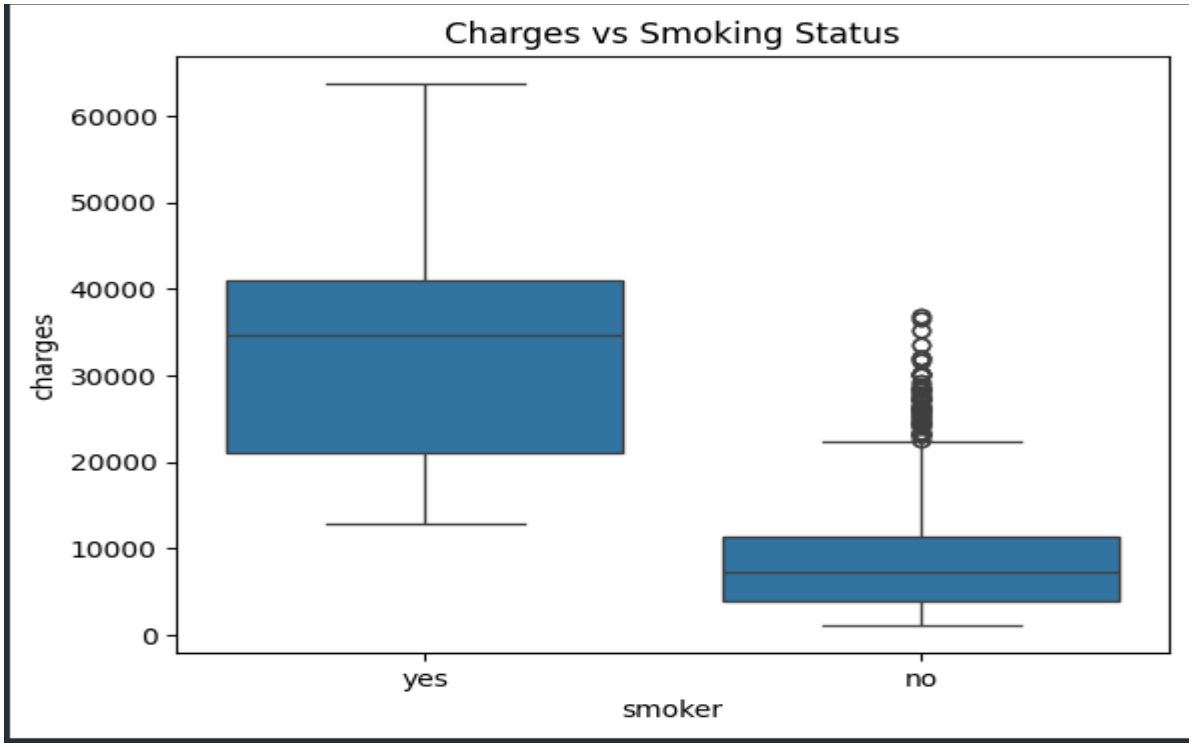


V. Result and Output

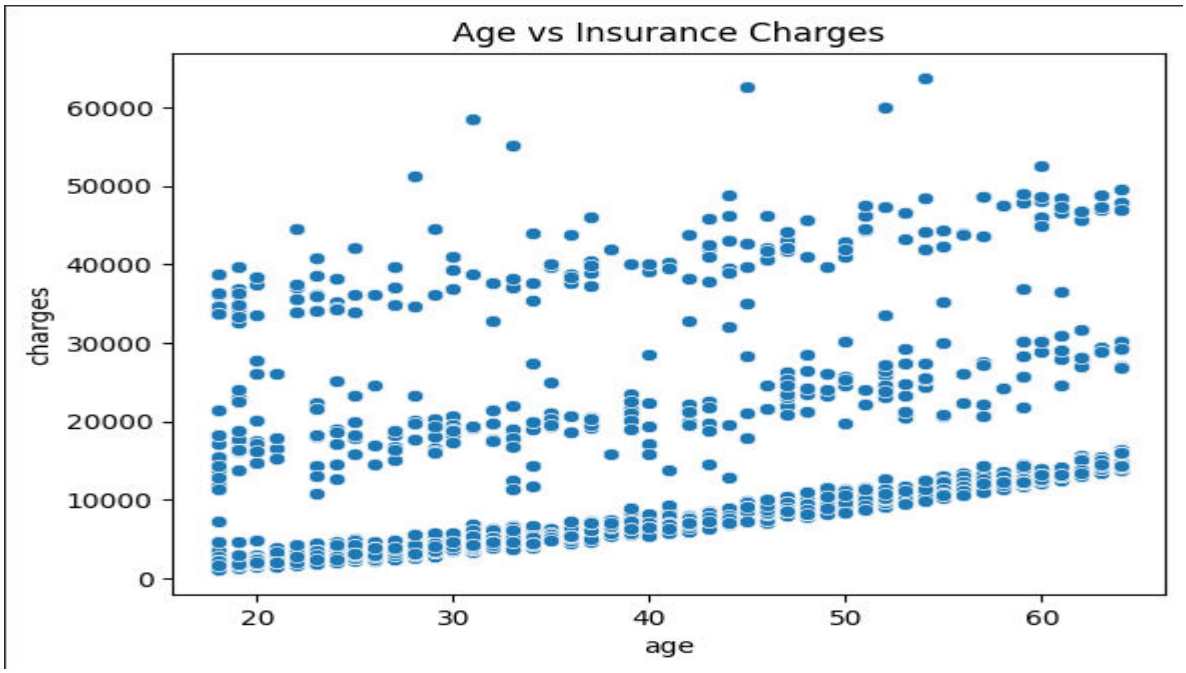
Distribution of Insurance Charges



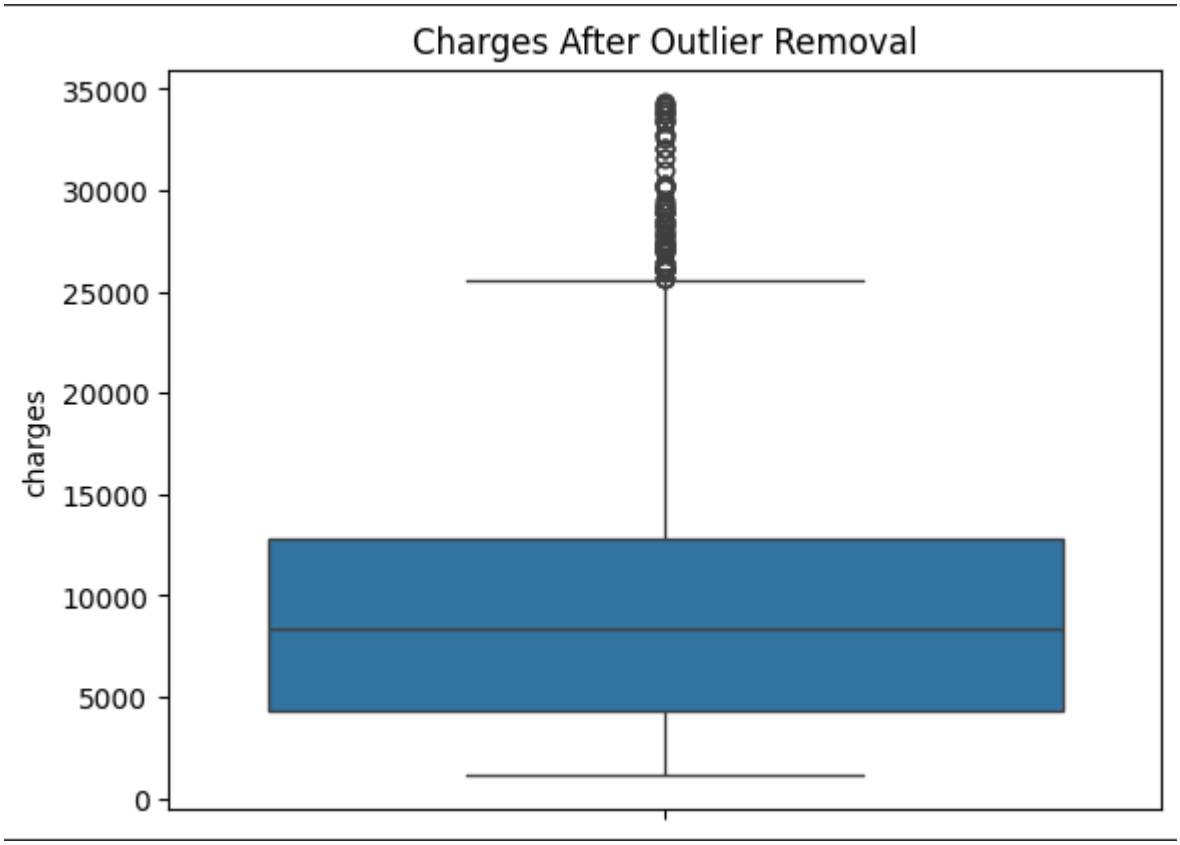
Charges vs. Smoking Status



Age vs. Insurance Charges



Charges After Outlier Removal



Model Performance-Random Forest Regressor

Model Evaluation Results

Medical Insurance Cost Prediction Performance

Metric	Result
Mean Absolute Error (MAE)	1242.28 Average error of 1242 units
Root Mean Squared Error (RMSE)	2403.51 Measures larger prediction errors
R^2 Score	0.885 Explains 88.5% of variance
Training Score	0.979 High accuracy on training set
Test Score	0.885 Good performance on test set

VI. Conclusion

This project demonstrates the effectiveness of machine learning techniques in predicting medical insurance costs with high accuracy. The Random Forest model, achieving an R^2 score of 0.885, proves to be a reliable approach for capturing complex relationships between various factors and insurance charges. Key features such as smoking status, age, and BMI are identified as the most influential contributors to cost variation.

The developed model enables fairer pricing, improved risk assessment, and data-driven decision-making for insurance providers and policymakers. By leveraging predictive analytics, the system enhances transparency and efficiency in the insurance domain. Overall, this project highlights the transformative potential of machine learning in healthcare finance and paves the way for more advanced, accurate, and equitable insurance prediction systems in the future.

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.