International Journal of Innovative Research in Engineering
Volume 6, Issue 2(March-April 2025), PP: 41-43.
www.theijire.com **ISSN No: 2582-8746**

Insurance Cost Prediction Using Polynomial Ridge Regression and Random Forest Classifier

G. Rajeswari[1], J.V. Kowsalya[2,] T.S. Navalakshmi[3], N.S Sakthi Aashitha[4], T.S Sakthi Jothi[5]
[1]Assistant Professor, Department of Computer Science and Engineering, K.L.N. College of Engineering, Sivagangai, Tamilnadu, India.
[2,3,4,5] Final Year Students, Department of Computer Science and Engineering, K.L.N. College of Engineering, Sivagangai, Tamilnadu, India.

How to cite this paper:
G. Rajeswari[1], J.V. Kowsalya[2], T.S. Navalakshmi[3], N.S Sakthi Aashitha[4], T.S Sakthi Jothi[5] "'Insurance Cost Prediction Using Polynomial Ridge Regression and Random Forest Classifier", IJIRE-V6I02-41-43.

Abstract: *Machine Learning, particularly Polynomial Ridge Regression and Random Forest Classifier, is used to predict medical insurance costs with accuracy and personalization. Polynomial Ridge Regression models complex relationships for cost estimation, while Random Forest Classifier categorizes individuals into cost groups. Key factors like age, BMI, smoking, and region serve as inputs. Data preprocessing ensures accuracy, while real-time feedback enhances predictions. Feature importance analysis helps assess financial risks. With scalability and a user-friendly interface, the system supports transparent, reliable, and adaptive insurance cost management.*

Key Word: *Machine Learning, Polynomial Ridge Regression, Random Forest Classifier, Healthcare Cost Variations.*

I.INTRODUCTION

The Medical Insurance Cost Prediction System enhances traditional cost estimation by leveraging machine learning for greater accuracy and personalization. Using Polynomial Ridge Regression, the system captures complex relationships between factors such as age, BMI, smoking habits, number of children, and region to predict insurance premiums. Meanwhile, the Random Forest Classifier categorizes individuals into Low, Medium, and High-cost groups, providing clear risk assessment. By automating data preprocessing and integrating real-time feedback, the system continuously improves its predictions. This advanced approach helps insurers design fair premium structures while enabling policyholders to make informed financial decisions.

II. OBJECTIVE

Machine Learning, particularly Polynomial Ridge Regression and Random Forest Classifier, is used in this study to predict medical insurance costs accurately and transparently. Polynomial Ridge Regression models complex relationships for precise cost estimation, while Random Forest Classifier categorizes individuals into Low, Medium, and High-cost groups. Key factors such as age, BMI, smoking habits, number of children, and region serve as input variables. Data preprocessing ensures accuracy, and real-time feedback enhances predictions. This dual-model approach helps insurers design fair premium structures while enabling policyholders to make informed financial decisions about their healthcare coverage.

III. LITERAURE SURVEY

1. Current methods of ROCK vs MINE Prediction:

Traditional insurance cost estimation methods relied on generalized assumptions, often leading to inaccurate and non-personalized predictions. These approaches lacked the ability to capture complex relationships between factors such as age, BMI, smoking habits, number of children, and region, resulting in inconsistencies in premium calculations. Additionally, these methods did not effectively address data imbalances, reducing their reliability and fairness in cost estimation.

2. Technology's role in improving quality:

Medical insurance cost prediction is enhanced using machine learning, specifically Polynomial Ridge Regression for precise cost estimation and Random Forest Classifier for categorizing individuals into Low, Medium, and High-cost groups. To handle data imbalances, under-sampling and oversampling techniques improve accuracy. The model is evaluated by using accuracy, precision, recall, and F1-score. Data visualization tools in machine learning such as heatmaps and ROC curves, provide insights for insurers, enabling fair premium structures and helping policyholders make informed financial decisions.

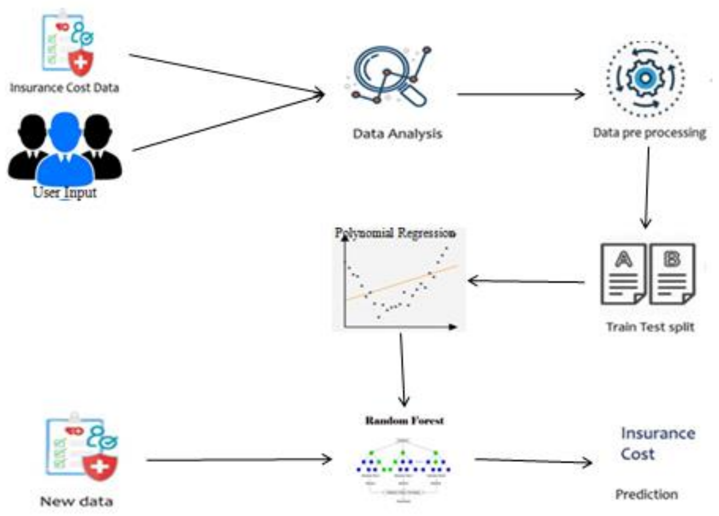
IV. EXISTING SYSTEM

Predictive modeling in healthcare leverages machine learning methods like XGBoost, GBM, and Random Forest to achieve accurate medical cost predictions. Ensemble models effectively capture complex data patterns, providing reliable outcomes but requiring significant computational resources and fine-tuning. While tools like SHAP help explain model decisions, their complexity often limits interpretability for end-users. Additionally, scalability and operational integration remain challenges due to the resource demands and complexity of ensemble models.

V. PROPOSED SYSTEM

Unlike traditional systems, this model integrates factors like age, BMI, smoking, and region for personalized insurance cost predictions. Polynomial Ridge Regression ensures precise cost estimation, while Random Forest Regressor enhances accuracy. Real-time updates optimize predictions, considering regional differences and improving user trust.

VI. ARCHITECTURE DIAGRAM



VII. SYSTEM OVERVIEW

1. Data Preprocessing and Manipulation

The data is collected, cleaned, and transformed to ensure accuracy. Missing values are handled through removal or imputation, while categorical data is encoded numerically. Numerical features are normalized, and the dataset is split into training and testing sets for effective model evaluation.

2. Exploratory Data Analysis and Visualization

The data is analyzed through visualizations to understand its structure. Numerical features like age, BMI, and charges are explored using histograms and box plots, while categorical features are examined with count plots. Correlation matrices and scatter plots reveal relationships, helping identify outliers and trends to improve model performance.

3. Model Building and Evaluation

The model is trained using Polynomial Ridge Regression and Random Forest Classifier to capture relationships between input features and the target variable. Performance is evaluated using R^2, MSE for regression, and accuracy or F1-score for classification. Feature importance and polynomial coefficients are analyzed to understand their impact, while residuals and confusion matrices ensure reliability. Hyper-parameters are fine-tuned, cross-validation is applied, and regularization techniques are used to enhance performance and prevent over-fitting.

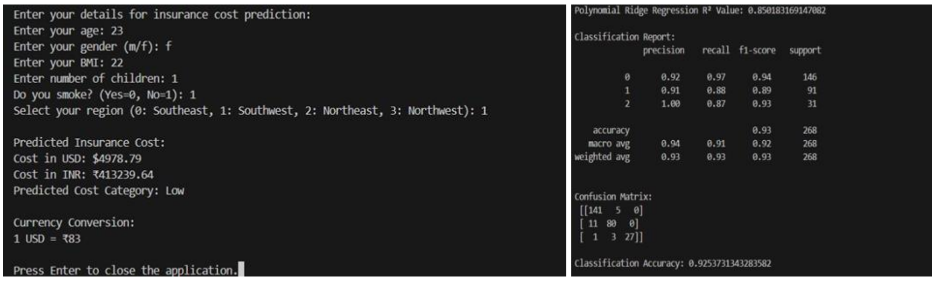
4. Prediction and Input Transformation

The system processes new input data by applying transformations, including scaling for Polynomial Ridge Regression and encoding for Random Forest Classifier. Trained models generate insurance cost estimates and classify individuals into cost brackets. Numerical and categorical outputs are translated into meaningful insights. Missing or new features are managed using imputation or default values to ensure consistency. Models are periodically retrained and fine-tuned with new data to improve accuracy and adapt to evolving patterns.

5. Monitoring and Optimization

The system continuously monitors model performance using metrics like R^2, MSE, accuracy, and F1-score to ensure reliable predictions. Discrepancies between predictions and actual insurance costs help identify data drift, model degradation, or shifts in data distribution. Regular assessments confirm the model meets accuracy standards, allowing timely improvements. Predictions are evaluated against real-world outcomes to maintain relevance in cost estimation. Automated monitoring tools detect anomalies and trigger updates, ensuring efficient performance and adaptability.

.



```
Enter your details for insurance cost prediction:
Enter your age: 23
Enter your gender (m/f): f
Enter your BMI: 22
Enter number of children: 1
Do you smoke? (Yes=0, No=1): 1
Select your region (0: Southeast, 1: Southwest, 2: Northeast, 3: Northwest): 1

Predicted Insurance Cost:
Cost in USD: $4978.79
Cost in INR: ₹413239.64
Predicted Cost Category: Low

Currency Conversion:
1 USD = ₹83

Press Enter to close the application.
```

```
Polynomial Ridge Regression R² Value: 0.850183169147082

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.97      0.94       146
           1       0.91      0.88      0.89        91
           2       1.00      0.87      0.93        31

    accuracy                           0.93       268
   macro avg       0.94      0.91      0.92       268
weighted avg       0.93      0.93      0.93       268


Confusion Matrix:
 [[141   5   0]
 [ 11  80   0]
 [  1   3  27]]

Classification Accuracy: 0.9253731343283582
```

Fig 7.1 User Input and Performance Metrics

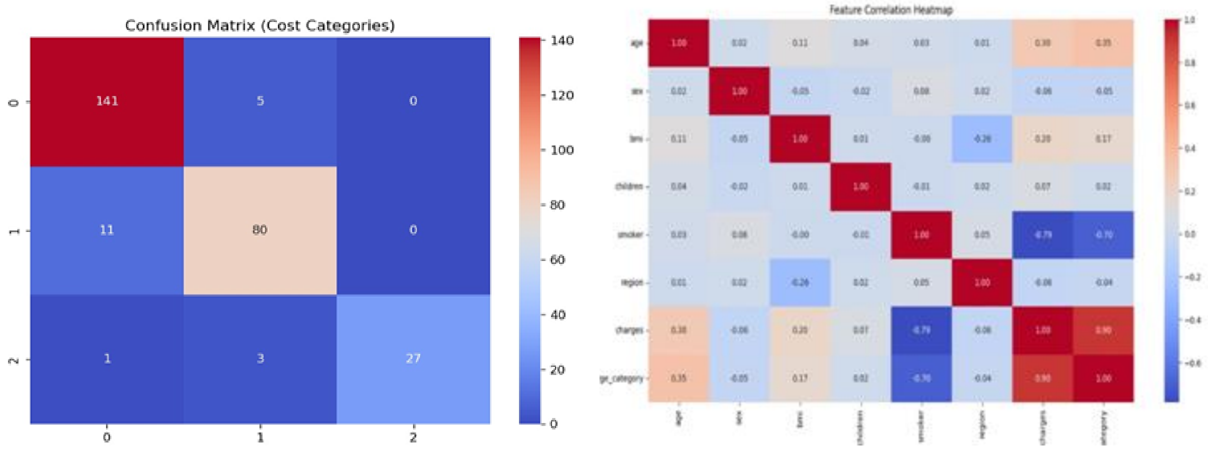


Fig 7.2 Pictorial Representation of Performance Metrics

VIII. CONCLUSION

- **Practical Machine Learning Implementation:** This study applies machine learning, specifically polynomial ridge regression, to classify user data, effectively predicting the medical insurance costs.
- **Systematic Workflow for Accuracy:** A structured approach, including data collection, preprocessing, model training, and evaluation, ensures reliable object classification.
- **Strong Classification Performance:** The Random forest regressor model achieved high accuracy, demonstrating its effectiveness in structured classification tasks.
- **Essential Preprocessing Techniques:** Feature scaling, label encoding, and proper data splitting played a key role in building a robust and accurate model.
- **Potential for Broader Applications:** The model's structured workflow and strong performance indicate its applicability to similar classification tasks requiring structured data analysis.

References

1. *A. Ravishankar Rao, Subrata Gardai, Coumarate Dey, Hang Peng, "Building predictive models of healthcare costs with open healthcare data",2020 IEEE International Conference on Healthcare Informatics (ICHI) | 978-1-7281-5382- 7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHI48887.2020.9374348*
2. *Sheng Yao Zhou, Run tong Zhang*, "A Novel Method for Mining Abnormal Expenses in Social Medical Insurance" Auckland University of Technology. Downloaded on November 07, 2020 at 17:01:50 UTC from IEEE Xplore.*
3. *Anuja Tike, Sanket Tavarageri, "A Medical Price Prediction System using Hierarchical Decision Trees",2017 IEEE International Conference on Big Data (BIGDATA)*
4. *Pei Shen, "Factor Analysis of Medical Expenses of the Hepatitis A patients in Guangdong", 2016 8th International Conference on Information Technology in Medicine and Education.*