**Review-3 Of BDA**

**Health Care Cost Analysis and Prediction**

**Team Members:**

**2210080009-I.Geeta Sai Sahasra**

**2210080020-K.Mahi Manusha**

**2210080021-B.Chinmayee Reddy**

**2210080065-P.Hema Sankhar**

**Abstract:** Health care cost analysis and prediction play a crucial role in the healthcare industry by enabling better financial planning, cost management, and patient affordability. This study aims to analyze healthcare costs using historical data and predict future expenses using machine learning models. Our implementation focuses on utilizing machine learning techniques to enhance predictive accuracy while addressing key challenges in healthcare cost forecasting.

**1. Introduction** The increasing complexity of healthcare expenses necessitates the need for robust analytical frameworks to understand and predict future costs. By leveraging machine learning models, healthcare providers and policymakers can gain insights into cost trends, helping to allocate resources efficiently. This project presents an approach to analyzing and predicting healthcare costs using machine learning methodologies.

**2. Objectives**

* To analyze historical healthcare cost data.
* To identify factors affecting healthcare expenses.
* To implement machine learning models for accurate cost prediction.
* To evaluate and improve prediction accuracy.

**3. Methodology** Our approach consists of multiple stages, including data collection, preprocessing, model training, and evaluation. Below is a structured breakdown of our methodology:

**3.1 Data Collection**

We utilize a dataset containing historical healthcare cost records, which includes factors such as patient demographics, medical conditions, treatment details, and hospital charges.

**3.2 Data Preprocessing**

* Handling missing values using mean/mode imputation.
* Normalizing and scaling numerical attributes.
* Encoding categorical variables (e.g., gender, diagnosis codes).
* Splitting data into training and testing sets.

**3.3 Model Selection**

We experimented with the following machine learning models:

* Linear Regression
* Decision Trees
* Random Forest
* Gradient Boosting Machines (GBM)
* Neural Networks

**3.4 Model Training**

We trained different models using the available dataset and evaluated their performance using key metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared scores.

**3.5 Model Evaluation**

After training, we assessed model accuracy and refined hyperparameters to optimize predictions. Feature importance analysis was conducted to understand key drivers of healthcare costs.

**4. Results and Findings** Initial results showed that:

* Linear Regression provided baseline predictions but struggled with complex relationships.
* Decision Trees captured non-linear patterns but overfitted on smaller datasets.
* Random Forest and GBM performed well, with lower error margins.
* Neural Networks showed promise but required further tuning for better generalization.

**Accuracy Improvements Needed:**

* Further dataset training required to improve prediction robustness.
* Hyperparameter optimization for neural networks.
* Feature engineering to incorporate additional cost factors.
* Ensemble techniques to combine model strengths for improved accuracy.

**5. Challenges and Future Work**

* **Data Limitations:** Additional data collection needed to enhance model generalization.
* **Feature Selection:** Identifying more relevant variables that influence healthcare costs.
* **Computational Complexity:** Advanced models like deep learning require high computational power.
* **Regulatory Considerations:** Ensuring compliance with healthcare data privacy laws.

Future improvements will focus on refining model accuracy, optimizing computational efficiency, and incorporating real-time cost prediction mechanisms.

**6. Review-3 Updates** For Review-3, we have:

* Developed the initial code structure for data preprocessing and model training.
* Conducted preliminary training but still require additional dataset refinement.
* Observed accuracy gaps that need tuning via hyperparameter optimization.
* Planned enhancements in feature selection and ensemble learning for better results.
* Scheduled further testing and evaluation to finalize the predictive model.

**7. Conclusion**

This project demonstrates the potential of machine learning in predicting healthcare costs. While initial results are promising, further refinements are needed to achieve higher accuracy. The ongoing enhancements will aim to deliver a more robust and reliable cost prediction framework.