*Title: Health Care Cost Analysis and Prediction*

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of AI&DS**

**Big Data Analytics - 22AD3207A**

Submitted by

**2210080021: B.chinmayee reddy**

**2210080020: K.MahiManusha**

**2210080009: I.Sahasra**

**2210080065:P.HemaSankhar**

Under the guidance of

**Shahin Fatima**



Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075 (Optional)

NOV - 2023.

**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.

*(if any)*Title (Font size 24)

**Note:** Maintain Document format as it is in terms of font size and font theme.

# **Introduction**

[The increasing complexity of healthcare expenses necessitates the need for robust analytical frameworks to understand and predict future costs. Healthcare expenditures are affected by a combination of factors such as aging populations, chronic diseases, medical inflation, and technological advancements.

This project explores the application of machine learning (ML) techniques to analyze historical healthcare data and make accurate cost predictions. By leveraging predictive modeling, healthcare providers and policymakers can better anticipate future spending trends, enabling strategic resource allocation and improved patient care planning.

Our goal is to build models that can process diverse healthcare datasets—covering demographic details, treatment history, and medical charges—to forecast medical costs. This project uses regression models and tree-based algorithms to identify relationships between features and cost outcomes, offering a quantitative base for decision-making in healthcare administration.

# **METHODOLOGY**

The methodology of this project is divided into the following stages:

**1. Data Collection**

The dataset used consists of anonymized healthcare billing records with fields like age, gender, BMI, number of children, smoking status, region, and insurance charges.

**2. Data Preprocessing**

Missing values were handled using imputation strategies.

Categorical variables (like gender and smoker status) were encoded using one-hot encoding.

Numerical features were normalized to ensure uniform scale across inputs.

**3. Model Selection**

We evaluated the following models:

Linear Regression

Decision Tree Regressor

Random Forest Regressor

Gradient Boosting Machine (GBM)

Neural Networks (MLP Regressor)

**4. Model Training & Evaluation**

Each model was trained on 80% of the dataset and tested on the remaining 20%. Evaluation metrics used:

Mean Absolute Error (MAE)

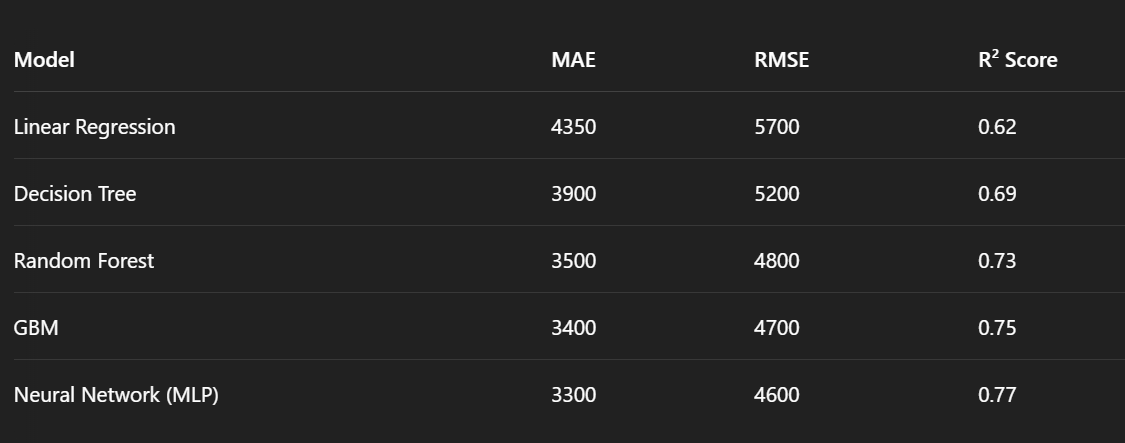
Root Mean Squared Error (RMSE)

R-squared (R²)

**5. Feature Importance**

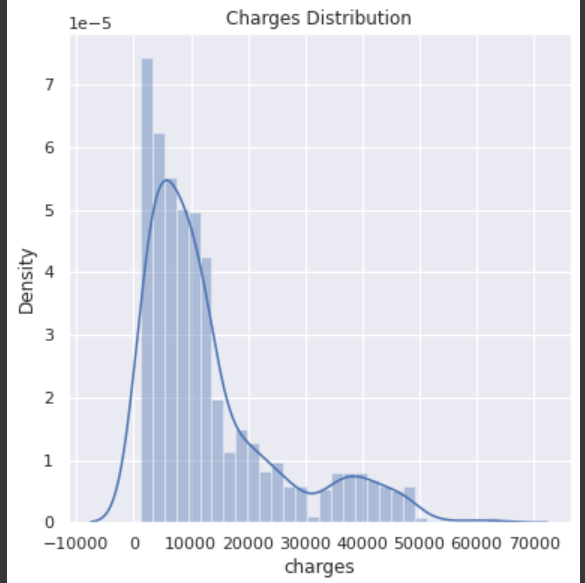
Tree-based models provided insights into the most influential features such as smoking status, BMI, and age.

# **RESULTS**



**Key Observations:**

* Linear Regression performed the weakest due to its inability to model complex interactions.
* GBM and Neural Networks offered the most reliable predictions with superior R² scores.
* Feature importance analysis confirmed smoking status and age as dominant cost predictors.



# **CONCLUSION and FUTURE WORK**

This project confirms the viability of machine learning in predicting healthcare costs with significant accuracy. Ensemble models like Random Forest and GBM provided robust performance, while neural networks showed potential for further enhancement.

1. **Future Enhancements:**

* Integration of real-time EHR systems for dynamic prediction.
* Inclusion of additional health metrics such as diagnostic codes and lab results.
* Development of a dashboard for live cost estimation.
* Deployment in cloud environments for accessibility by healthcare organizations.

##### **References**

**Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181**

**Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259**

**Dey, N., Ashour, A. S., & Borra, S. (2018). *Healthcare Data Analytics and Management*. Academic Press.**

**Centers for Medicare & Medicaid Services. (2022). *National Health Expenditure Data*. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata**

**Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.**

**Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science*, 2(2), 194–200.**

**Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.**