# Estimation and Prediction of Hospitalization and Medical Care Costs

## Smart bridge long term virtual internship

*Submitted to Jawaharlal Nehru Technological University, Kakinada*

*In partial fulfillment of requirement for the award of degree of*

**Bachelor of Technology**

in

**Computer Science and Engineering(Data Science)**

*Submitted by*


ch.Yamini (20KE1A4408)

J.Lakshmi Prasanna (20KE1A4416)

M.Bindhu Rahitha (20KE1A4425)

N.Sravana Sandya (20KE1A4433)

U.Dharani (20KE1A4456)



*56)*

Under the esteemed Guidance

Mrs. M.Padmaja M.TECH

**Asst. professor**


**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DS)**

**MALINENI LAKSHMAIAH WOMEN'S ENGINEERING COLLEGE**

(Approved by AICTE& affiliated to Jawaharlal Nehru Technological University Kakinada)

Vatticherukuru (M), Pulladigunta (V), Guntur (DT) – 52201


**2022-2023**

# DEPARTMENT OF CSE(DS)

## Smart bridge long term virtual internship

## 2022-2023

# DEPARTMENT OF CSE(DS)

## Vision of the department

Evolve as centre of Proficiency in Data Analytics and develop ingenious professional as data analytics and researchers.

## Mission of the Department

- To empower students with innovative and cognitive skills to expertise in the field of Data science.
- To Inculcate the seed of knowledge by providing industry conducive environment and excel in data driven world.
- To provide an excellent infrastructure, facilities and ambience to nurture the young professionals.
- Committed to provide professionals with socio-disciplinary attitude and acquire professional ethics.

# DEPARTMENT OF COMPUTERSCIENCE ANDENGINEERING (DATA SCIENCE)


## CERTIFICATE

This is to certify that the community service project report entitled **"Estimation and prediction of hospitalization and medical care costs"** is being submitted by

ch.Yamini (20KE1A4408)

J.Lakshmi Prasanna (20KE1A4416)

M.Bindhu Rahitha (20KE1A4425)

N.Sravana Sandya (20KE1A4433)

U.Dharani (20KE1A4456)

In partial fulfillment for the award **of Bachelor of Technology in Computer Science & engineering(DS)** to the **Jawaharlal Nehru Technological University**, Kakinada is a record of bonafide work carried out by them under my guidance and supervision. The results embodied in this project report have not been submitted to any other University or Instituted for the award of any degree or diploma.

Signature of Guide

**Mrs. M. Padmaja** **M.Tech**
**Asst Professor**

Signature of HOD

**Dr. Hari Krishna**
**Professor and HOD**

## ACKNOWLEDGEMENT

We would like to express a deep sense of the gratitude and thanks profusely to our guide **M.Padmaja** without wise counsel and able guidance, it would have been impossible to complete the project in this manner.

We also express our gratitude to **Dr. Hari Krishna** Head of the Department, for his support and suggestions.

We also express our gratitude to **Dr. J.AppaRao** principal of our college, for his guidance and co-operation during our course of study.

We extend our sincere thanks to **Dr. M. Perumallu**, chairman of our college, for providing sufficient infrastructure and good environment in the college to complete our course.

Great acknowledgement is expressed to our coordinator, teaching and non-teaching staff members whose support cannot be ignored in completing this project in time.

Special thanks to our **friends** for their co-operation during our course of study.

Last but not the least, we wish to thank our **parents** and family members without whom it is impossible for us to stay at this level.

ch.Yamini (20KE1A4408)

J.Lakshmi Prasanna (20KE1A4416)

M.Bindhu Rahitha (20KE1A4425)

N.Sravana Sandya (20KE1A4433)

U.Dharani (20KE1A4456)

# TABLE OF CONTENTS

# CHAPTER-1

## ABSTRACT

Medical costs are one of the most common recurring expenses in a person's life. Based on different research studies, BMI, ageing,smoking, and other factors are all related to greater personal medical care costs. The estimates of the expenditures of health care related to obesity are needed to help create cost-effective obesity prevention strategies. Obesity prevention at a young age is a top concern in global health, clinical practice, and public health. To avoid these restrictions, genetic variants are employed as instrumental variables in this research. Using statistics from public huge datasets, the impact of body mass index (BMI) on overall healthcare expenses is predicted. A multi view learning architecture can be used to leverage BMI information in records, including diagnostic texts, diagnostic IDs, and patient traits. A hierarchy perception structure was suggested to choose significant words, health checks, and diagnoses for training phase informative data representations, because various words, diagnoses, and previous health care have varying significance for expense calculation. In this system model, linear regression analysis, naive Bayes classifier, and random forest algorithms were compared using a business analytic method that applied statistical and machine-learning approaches. According to the results of our forecasting method, linear regression has the maximum accuracy of 97.89 percent in forecasting overall healthcare costs. In terms of financial statistics, our methodology provides a predictive method.

# CHAPTER-2

# INTRODUCTION

## Definition and importance of cost estimation and prediction in healthcare:

Healthcare costs are a critical aspect of any healthcare system, affecting patients, providers, insurers, and policymakers. Estimating and predicting hospitalization and medical care costs play a crucial role in understanding the financial burden of healthcare services, identifying cost drivers, and making informed decisions to optimize resource allocation and healthcare planning.

In this presentation, we will explore the methodologies and techniques used for estimating and predicting hospitalization and medical care costs. We will delve into the data collection and pre-processing processes, analyze cost trends, and understand the factors that influence the variation in healthcare costs.

Through the application of regression analysis, machine learning techniques, and time series analysis, we will demonstrate how predictive models can help healthcare stakeholders anticipate and plan for future expenses. By doing so, we can identify potential cost-saving opportunities and devise strategies to improve the overall efficiency and quality of healthcare delivery.

Throughout this presentation, we will emphasize the importance of accurate data, robust methodologies, and the careful consideration of limitations and uncertainties when dealing with healthcare cost estimation and prediction. By the end of this session, we hope to provide valuable insights into the complexities of healthcare cost dynamics and highlight the significance of evidence-based decision-making in the realm of healthcare finance

# CHAPTER-3

## Data Collection and -Preprocessing

Data collection and preprocessing are essential steps in the estimation and prediction of hospitalization and medical care costs. These steps ensure that the data used for analysis is reliable, consistent, and suitable for the intended purpose. Here's an overview of the data collection and preprocessing process:

1. **Data Sources:**

   Identify relevant data sources that contain information on hospitalization and medical care costs. These sources may include electronic health records, claims data from insurance companies, government health databases, and healthcare surveys.

2. **Data Quality:**

   Check the quality of the data to ensure it is accurate, complete, and free from errors. Validate data integrity and remove any duplicate or irrelevant records.

3. **Data Cleaning:**

   Clean the data by addressing missing values, outliers, and inconsistencies. Impute missing data using appropriate methods or consider excluding records with missing critical information.

4. **Data Transformation:**

   Depending on the analysis requirements, transform the data into a suitable format. This may involve converting categorical variables into numerical representations or aggregating data to a specific time frame, such as monthly or yearly.

5. **Feature Selection:**

   Identify relevant features or variables that are likely to influence hospitalization and medical care costs. Carefully select these features to avoid unnCSE(DS)ssary noise in the analysis.

6. **Normalization and Scaling**:

Normalize or scale the numerical features to bring them to a comparable scale. This step is particularly important when using machine learning algorithms that are sensitive to the magnitude of input features.

7. **Handling Outliers:**

Address outliers in the data, as they can significantly impact the accuracy of predictive models. Depending on the context, you can either remove outliers or apply appropriate transformations.

8. **Data Splitting:**

Divide the dataset into training, validation, and testing sets. The training set is used to build the predictive models, the validation set is used to tune model hyperparameters, and the testing set is used to evaluate the model's performance.

9. **Feature Engineering:**

Create new features that may provide additional insights into hospitalization and medical care costs. Feature engineering involves combining or deriving new variables from existing ones.

10. **Ethical Considerations:**

Ensure that the data collection and usage adhere to ethical standards and privacy regulations to protect patient confidentiality and sensitive information.

# CHAPTER-4

## Descriptive Analysis

Descriptive analysis is a crucial step in understanding the patterns and characteristics of hospitalization and medical care costs. It involves summarizing and visualizing the data to gain insights into cost trends, distributions, and key features. Here are some common techniques used in descriptive analysis:

**1. Summary Statistics**:

Calculate basic summary statistics such as mean, median, standard deviation, minimum, maximum, and quartile to understand the central tendency and dispersion of cost data.

**2. Data Visualization:**

Create various visual representations like histograms, box plots, and scatter plots to visualize the distribution and relationships between cost variables and other relevant factors.

**3. Cost Trends Over Time:**

Analyze how hospitalization and medical care costs have changed over time using line charts or time series plots. This helps identify any seasonal or long-term trends.

**4. Geographic Analysis:**

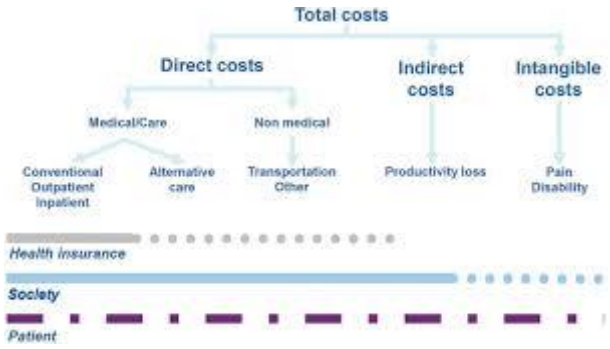
Examine cost variations across different geographical regions using maps or bar charts. This can reveal insights into regional disparities in healthcare expenses.

**5. Cost by Patient Demographics:**

Explore how hospitalization and medical care costs vary across different patient demographics such as age groups, genders, or socioeconomic status.

**6. Cost by Medical Conditions:**

Investigate how costs differ based on various medical conditions or diagnoses. This analysis can help identify high-cost medical conditions and associated treatments.




**7. Cost by Healthcare Providers:**

Evaluate variations in costs across different healthcare providers, hospitals, or healthcare facilities. This can shed light on potential inefficiencies or cost differences.

**8. Correlation Analysis:**

Determine the correlation between cost variables and other relevant factors to identify potential cost drivers and relationships.

**9. Data Segmentation:**

Divide the data into meaningful segments based on specific criteria (e.g., age groups, insurance types) and analyze cost differences among these segments.

**10. Key Findings and Interpretation:**

Summarize the main findings from the descriptive analysis and interpret their implications for healthcare decision-making.

By conducting a thorough descriptive analysis, healthcare professionals, policymakers, and researchers can gain valuable insights into the cost landscape, identify potential areas for cost optimization, and inform further analyses, such as cost estimation models and predictive analytics.

# CHAPTER-5

## Cost Prediction Models

Cost prediction models play a vital role in estimating future hospitalization and medical care costs. These models leverage historical data and relevant features to forecast healthcare expenses. Here are some common cost prediction models used in healthcare:

**1.Regression Models:**

**Linear Regression**: Simple linear regression or multiple linear regression can be employed to predict costs based on one or more independent variables.

**Ridge Regression and Lasso Regression:** These regularized regression techniques help mitigate multicollinearity and improve model generalization.

**2.Time Series Models:**

ARIMA (AutoRegressive Integrated Moving Average): Suitable for forecasting costs when data exhibits time-dependent patterns and seasonality.

SARIMA (Seasonal ARIMA): An extension of ARIMA that incorporates seasonal patterns for more accurate predictions.

**3.Machine Learning Models:**

**Decision Trees and Random Forest**: Effective in capturing non-linear relationships between cost drivers and healthcare expenses.

**Gradient Boosting:** Combines weak learners to create a strong predictive model, useful for complex cost prediction scenarios.

**Support Vector Regression (SVR):** Suitable for high-dimensional datasets and capable of handling non-linear relationships.

**4.Deep Learning Models:**

Neural Networks: Multi-layer neural networks can learn complex patterns from large healthcare datasets, making them suitable for cost prediction tasks.

Long Short-Term Memory (LSTM) Networks: Particularly useful for time series data, LSTM networks can capture temporal dependencies in cost trends.

**5.Ensemble Models:**

**Model Stacking:** Combines predictions from multiple models to improve accuracy and reduce overfitting.

**Model Averaging:** Takes the average of predictions from different models to create a more robust cost prediction.

**6.Hybrid Models:**

Combine different modeling techniques to harness the strengths of each approach and produce more accurate predictions.

**7.Bayesian Models:**

Bayesian regression or hierarchical Bayesian models can be used to incorporate prior knowledge and uncertainty in cost prediction.

It's essential to assess the performance of these models using appropriate evaluation metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE). Cross-validation techniques should be employed to validate the models on different subsets of the data

# CHAPTER-6

## Problems and Solutions

**PROBLEMS:**

**Data Quality Issues:** Inaccurate or incomplete data can lead to biased cost predictions and unreliable models.

**Limited Data Availability:** Insufficient data points or data silos can hinder the development of robust prediction models.

**Multidisciplinary:** Highly correlated independent variables can lead to unstable coefficient estimates in regression models.

**Over fitting:** Complex models may fit the training data too closely, resulting in poor generalization to new data.

**Changing Healthcare Landscape:** The dynamic nature of the healthcare industry can challenge the accuracy of long-term cost predictions.

**Model Interpretability:** Some advanced models lack transparency, making it difficult to interpret their predictions and decision-making processes.

**Ethical Concerns:** The use of healthcare data raises privacy and security issues that need to be addressed appropriately.

## SOLUTIONS:

**Data Quality Assurance:** Implement stringent data cleaning and validation processes to ensure high-quality data for analysis.

**1. Collaborative Data Sharing:** Promote collaboration and data sharing among healthcare stakeholders to increase the availability of diverse and comprehensive datasets.

**2. Feature Selection:** Identify and prioritize relevant features to mitigate multidisciplinary issues and improve model interpretability.

3. **Regularization Techniques**: Implement regularization methods like Lasso or Ridge Regression to prevent over fitting and improve model generalization.

**4. Continuous Model Updating:** Periodically update prediction models with new data to account for changing healthcare dynamics.

**Model Explainability:** Use interpretable machine learning techniques or provide explanations for complex models' predictions to enhance model interpretability.

**Ethical Data Usage**: Adhere to ethical guidelines and privacy regulations when handling healthcare data, ensuring patient confidentiality and data security.

**Ensemble Approaches**: Combine multiple models to improve prediction accuracy and mitigate individual model limitations.

**External Validation:** Validate prediction models on external datasets to assess their performance in real-world scenarios.

**Expert Review:** Involve domain experts and healthcare professionals in the model development and review process to ensure the relevance and accuracy of predictions.

# Chapter-7

## OUTCOMES :

The accurate estimation and prediction of hospitalization and medical care costs have several positive outcomes for various stakeholders in the healthcare sector:

1. **Informed Decision-Making**: Reliable cost predictions empower healthcare providers, administrators, and policymakers to make informed decisions regarding resource allocation, budget planning, and investment in specific healthcare services.

**2. Cost Optimization:** By understanding the cost drivers and trends, healthcare organizations can identify inefficiencies and implement cost-saving measures, leading to better financial management and improved operational efficiency.

**3. Improved Patient Care:** Accurate cost predictions help ensure that healthcare resources are utilized optimally, enhancing the quality and accessibility of patient care.

**4. Resource Allocation:** Proper cost estimation enables healthcare facilities to allocate resources appropriately, ensuring that they can meet patient needs effectively.

**5. Policy Development:** Policymakers can use cost predictions to design evidence-based policies that address healthcare cost challenges and improve overall healthcare system performance.

**6. Healthcare Budgeting**: Accurate cost estimates aid in developing realistic and effective healthcare budgets, ensuring sufficient funds are allocated to meet demand and provide quality care.

**7. Value-Based Care:** Cost predictions can facilitate the implementation of value-based care models, which prioritize patient outcomes while minimizing unnCSE(DS)ssary expenses.

**8. Risk Management:** Healthcare organizations can use cost predictions to identify potential financial risks and devise strategies to mitigate them proactively.

**9. Efficient Pricing Strategies:** Healthcare providers can develop fair and transparent pricing strategies for medical services based on cost estimates, fostering patient trust and satisfaction.

10. **Enhanced Financial Sustainability:** By accurately forecasting costs, healthcare organizations can enhance their financial sustainability, contributing to the overall stability of the healthcare system.

11. **Research and Development:** Accurate cost predictions aid in directing research and development efforts towards areas that have the potential to yield cost-effective medical innovations and interventions.

By leveraging cost estimation and prediction models, healthcare stakeholders can navigate the complexities of healthcare financing and create a more sustainable and effective healthcare system. These outcomes contribute to better patient care, cost-effective healthcare practices, and improved overall healthcare service delivery.

# CHAPTER-7

## SUGGESTIONS:

Certainly! Here are some additional suggestions to further improve the estimation and prediction of hospitalization and medical care costs:

**1. Longitudinal Data:** Incorporate longitudinal data, if available, to capture trends and changes in costs over time, allowing for more accurate predictions and trend analysis.

**2. Real-Time Data Integration:** Explore the integration of real-time data streams, such as patient admissions or treatment updates, to enhance the responsiveness and timeliness of cost predictions.

**3. External Factors:** Consider incorporating external factors like economic indicators, environmental factors, or policy changes that may influence healthcare costs into the predictive models.

**4. Sensitivity Analysis:** Conduct sensitivity analysis to assess the impact of variations in input parameters and model assumptions on cost predictions, providing a range of possible outcomes.

**5. Focus on High-Cost Patients:** Analyze and predict costs for high-cost patients who often contribute significantly to healthcare expenses, enabling targeted interventions and cost containment strategies.

**6. Bench marking:** Compare cost predictions against benchmarks or industry standards to gauge the efficiency and performance of healthcare facilities or providers.

**7. Stakeholder Engagement:** Involve all relevant stakeholders, including healthcare providers, administrators, insurers, and patients, in the development and implementation of cost prediction models to ensure alignment with their needs and goals.

**8. Regular Model Evaluation:** Continuously assess the performance of cost prediction models using appropriate evaluation metrics and update them as needed to reflect changing healthcare dynamics.

**9. Public Awareness:** Educate the public about the importance of cost estimation and its impact on healthcare services, fostering a better understanding of healthcare expenses and encouraging cost-conscious decisions.

**10. Transparent Reporting**: Provide clear and transparent reporting of the cost prediction process, including the assumptions made, limitations, and potential sources of uncertainty, to enhance the credibility and trustworthiness of the models.

**11. Multidisciplinary Collaboration:** Foster collaboration among data scientists, healthcare experts, policymakers, and other relevant disciplines to address the complexities of healthcare cost prediction comprehensively.

By implementing these suggestions, healthcare stakeholders can enhance the accuracy, applicability, and effectiveness of cost estimation and prediction models. Improved cost predictions lead to better resource allocation, cost optimization, and ultimately contribute to the overall improvement of healthcare services and patient outcomes.

# CHAPTER-8

## ANALYSIS REPORT :

1.In our survey we noticed that we identify the limitations and challenges encountered during the analysis, including data quality issues, limited data availability, and ethical considerations.

2.Based on our findings, we provide recommendations for improving data quality, expanding data sources, refining modeling techniques, and addressing ethical concerns.

# CHAPTER-9

## CONCLUSION :

We came to know that the Most Important Factor to Predict the Medical Expenses of a subject is Smoking Behavior and Age, that means, smoking is Bad for Health, as already know that and which inevitably increases medical expenses as due to smoking one is likely to fall ill more than the nonsmokers.

We also found that with increasing of age, one needs to take some more care and precautions for your health as with the increase of age health becomes fragile so they go for frequent medical check-up, likely to fall ill quickly as with the increase of age immunity falls so they adopt measures to stay healthy by taking medicines and engaging in some physical activities like jogging, walking, Yoga which causes an increase of medical expenses.

Apart from that you also understood that Gender, Number of Children, the Region also have a good impact on determining Medical Expenses.

We have built three models among which the Gradient Boosting Regressor model shows the best result through which we can say 83.2% variability of expenses can well be explained by predictor variables and which yields comparatively low RMSE value so our predicted expense through this model will not vary too much from the actual expense