FACTORS AFFECTING HEALTHCARE PRICING

Authors:

Savli Palande(A008), Yashasvi Mahadik(A036), Nidhi Joshi(A040)

Affiliation:

Nilkamal School of Mathematics, Applied Statistics and Analytics, SVKM's NMIMS, Mumbai

Mentor/Guide: Prof. Dr. Yogesh Naik

Abstract

Healthcare insurance pricing is a complex process influenced by various demographic, medical, and behavioural factors. Insurers analyse attributes such as age, BMI, smoking status, geographic location, and medical history to assess risk and determine premium costs. This research leverages data science techniques to identify key determinants affecting insurance pricing and develop predictive models for accurate premium estimation. By applying machine learning algorithms such as Linear Regression, Random Forest, and XGBoost on a publicly available dataset, we aim to enhance risk assessment methodologies and pricing fairness. Our findings indicate that smoking status and BMI significantly impact premium costs, with XGBoost achieving the highest predictive accuracy (R^2 = 86.81%). The study also highlights challenges such as data limitations, lack of real-time pricing models, and unexplored behavioural influences. These insights can assist insurers in optimizing pricing strategies while ensuring fairness and transparency in healthcare coverage. Future research should incorporate larger datasets, real-time health monitoring, and advanced AI techniques for further model refinement. The **Healthcare Pricing Prediction App** is a machine learning-based tool designed to estimate medical insurance costs based on key individual factors such as age, BMI, number of children, sex, smoking status, and region. By analyzing historical healthcare data, the application provides accurate and personalized predictions, assisting individuals, insurance companies, and healthcare providers in financial planning. The app features a user-friendly interface, ensuring accessibility for a wide range of users. With data-driven insights and statistical modeling, this tool enhances decision-making in healthcare expenditure estimation, offering a reliable approach to understanding and forecasting medical costs.

Keywords: Healthcare Insurance, Pricing Factors, Machine Learning, Predictive Modelling, Risk Assessment, Premium Estimation

1. Introduction

Health insurance plays a fundamental role in ensuring financial security and access to healthcare services for individuals and families. However, the pricing of health insurance premiums is a complex process influenced by multiple factors, including demographic characteristics, medical history, economic conditions, and policy regulations. Striking a balance between affordability, risk assessment, and profitability is essential to maintaining a sustainable health insurance system.

Traditional actuarial methods have long been used to calculate insurance premiums based on statistical risk models. However, these methods often fail to capture the complexities and dynamic nature of healthcare expenditures. In recent years, machine learning has emerged as a powerful tool in predictive analytics, enabling more accurate and efficient forecasting of insurance pricing. By leveraging large datasets and advanced computational techniques, machine learning models can identify patterns, trends, and risk factors that traditional models might overlook.

This study aims to explore the key determinants of health insurance pricing and assess the effectiveness of different predictive models in estimating insurance costs. The research compares multiple machine learning algorithms—Linear Regression, Decision Trees, Random Forest Regressor, and XGBoost—to determine the most accurate approach for price prediction. Among these, **Random Forest** demonstrated superior predictive performance, outperforming other models due to its ability to handle non-linear relationships and optimize error minimization.

The research objectives of this study include:

- Identifying the primary factors influencing
- health insurance pricing.
- Evaluating the strengths and weaknesses of different predictive modelling approaches.
- Assessing the impact of machine learning on improving pricing accuracy and fairness.

The findings of this study have significant implications for both insurance providers and consumers. Insurers can leverage predictive analytics to develop more precise premium structures, reducing inefficiencies and preventing unfair pricing practices. Additionally, consumers may benefit from increased transparency in insurance pricing, ensuring that premiums are determined based on objective risk assessments rather than opaque pricing models. Policymakers can also use these insights to regulate pricing frameworks, ensuring that machine learning models do not introduce biases or inequalities in healthcare accessibility.

This paper is structured into six sections. The first section provides a theoretical background on health insurance pricing models and key economic principles influencing premium calculations. The second section reviews existing literature, identifying major research themes and ongoing debates. The third section critically analyses gaps in the current research, highlighting areas that require further exploration. The fourth section presents an overview of methodologies used in predictive modelling, explaining the research process and evaluation techniques. The fifth section discusses empirical case studies that demonstrate real-world applications of machine learning in insurance pricing. Finally, the study concludes by summarizing key findings, discussing policy implications, and suggesting future research directions.

By integrating machine learning techniques into insurance pricing models, this research contributes to a growing body of work that seeks to enhance the accuracy, fairness, and efficiency of health insurance premium calculations. The study not only provides a comparative analysis of different modelling techniques but also offers insights. The **Healthcare Pricing Prediction App** is a powerful

tool designed to estimate medical insurance charges based on individual factors such as age, BMI, number of children, sex, smoking status, and region. By leveraging machine learning models, the app analyses user inputs to generate accurate predictions of healthcare costs. With a user-friendly interface, it allows individuals, insurance companies, and healthcare providers to gain valuable insights into expected medical expenses. The app provides personalized predictions based on real-world data, making it a reliable and efficient tool for financial planning in healthcare.

2. Literature Review

In recent years, health insurance coverage and costs have experienced a significant transformation in the United States. The Affordable Care Act, which is also known as Obamacare, is a prominent reform in order to address these challenges by establishing a Health Insurance Marketplace to promote available health insurance choices [2]. The study analysed data collected after the implementation of the ACA to assess the law's impact on health insurance affordability as well as individuals' ability to access care. The study looked at changes in insurance costs, out-of-pocket costs, and the total cost of medical services, as well as across different populations. This is the core background of our research. Choi and Blackburn explored patterns and factors in medical costs and health insurance premium payments

The research aims to identify various trends and factors that affect individual and household medical costs and health insurance costs, such as age, income level, etc. They also researched factors that contribute to discrepancies in the payment of medical and health insurance costs. These factors may include personal health status, medical services utilization, geographic location, and the insurance type that people hold. These factors are the reference for the source of our variable selection. The rise of algorithmic prediction and its implications were scrutinized by Cevolini and Esposito

They highlighted the transition in risk assessment from aggregation to analysis and discussed the social consequences of these algorithmic practices. The research also revealed potential bias and ethical considerations in predictive models. To complement these perspectives, Ch. Anwar ul Hassan et al. proposed that insurance companies are increasingly using algorithms and predictive analytics to assess risk, determine premiums, and make coverage decisions

Their research used machine learning models to predict insurance costs and compare the models' accuracy. The result showed that the Stochastic Gradient Boosting (SGB) model had the best performance. Kaushik et al. also used an artificial neural network (ANN) to predict the health insurance premium with an accuracy of 92.72%

However, according to Bhardwaj and Anand, the Gradient Boosting Regression Model was the best-performing model [7]. Through the synthesis of these studies, it is evident that the analysis of health insurance factors extends beyond traditional methods. The advent of advanced computing and machine learning techniques has paved the way for more accurate forecasts and a deeper understanding of dynamics in the health insurance space. We seek to build upon these foundations by using data visualization and regression models to explore the factors that would significantly a ect health insurance charges.

A brief report of the literature study is presented here. Only 17% of families in India covered any form of health insurance according to Mr. Shijith and Dr. T.V.Srkhar. Nevertheless, current health insurance statistics indicated the considerable increase of insured individuals and the number of health insurance plans during 2007-08. In 2008-09, the policy figures were 45, 75,725; in 2009-010, the policy figures grew to 68, 84,687 (TPA-served only). In metropolitan regions, higher coverage is recorded for health insurance. The coverage remains relatively low in rural regions.

3. Methodology

The following steps are followed for the methodology approach:

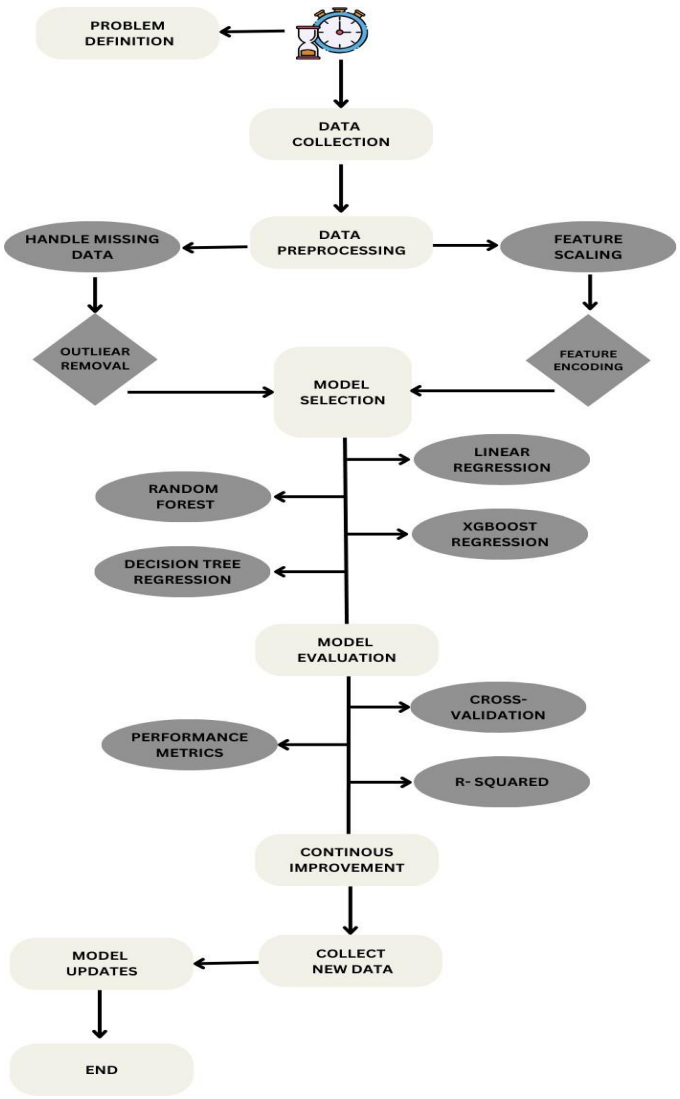


Fig 1

3.1 Data Collection

The dataset for this research was obtained from Kaggle and contains 1338 records. The key attributes considered for analysis include:

- Age: Older individuals generally have higher insurance costs due to increased health risks.
- BMI (Body Mass Index): A higher BMI is often associated with greater healthcare expenses.
- Smoking Status: Smokers tend to have significantly higher insurance premiums due to increased health risks.
- Number of Children: Family size can influence premium calculations.
- Region: Geographic location affects healthcare costs due to variations in medical services and regulations.
- Insurance Charges: This is the dependent variable representing the premium cost.

Feature Selection & Justification:

- **Age**

Why it matters? Older individuals are more prone to chronic illnesses and require more frequent medical care, leading to higher insurance premiums.

Statistical Basis: Healthcare expenses generally increase with age, particularly after 40-50 years, when risks of cardiovascular diseases, diabetes, and other conditions rise.

Model Consideration: The relationship between age and insurance charges may be non-linear (e.g., premiums increasing exponentially with age).

- **BMI (Body Mass Index)**

Why it matters? BMI is a key health risk indicator. Higher BMI is associated with obesity related diseases like diabetes, hypertension, and heart disease, which increase healthcare costs.

Statistical Basis: Studies show that individuals with a BMI > 30 (obese category) have significantly higher medical costs than those with a normal BMI.

Model Consideration: Instead of using BMI as a continuous variable, you can categorize it into:

- Underweight (<18.5)
- Normal (18.5 – 24.9)
- Overweight (25 – 29.9) Obese (≥30)

- **Smoking Status**

Why it matters? Smokers are at a higher risk of lung disease, heart conditions, and cancer, which increases their insurance premiums.

Statistical Basis: Smokers have 14 times higher medical costs related to lung diseases than non-smokers. Many insurance companies charge nearly double the premium for smokers compared to non-smokers.

Model Consideration: This is typically treated as a binary categorical variable (Yes/No).

- **Number of Children**

Why it matters? More children mean higher healthcare expenses, as they are covered under family health insurance plans.

Statistical Basis: Families with more dependents tend to have higher claim frequencies, increasing total insurance costs.

Model Consideration:

- Treated as a numerical variable (0, 1, 2, 3, etc.)
- You could also create a binary category (No children = 0, Has children = 1).

- **Region**

Why it matters? Healthcare costs vary based on geographic location due to differences in:

- State regulations on insurance
- Availability and cost of medical services
- Lifestyle and environmental factors

Statistical Basis: Urban areas tend to have higher insurance premiums due to expensive healthcare facilities. Some regions may have higher disease prevalence, affecting overall insurance pricing.

Model Consideration: Treated as a categorical variable (e.g., Northeast, Southeast, Southwest, Northwest). One-hot encoding can be applied for machine learning models.

- **Insurance Charges (Target Variable)**

Why it matters? This is the dependent variable (the value we are trying to predict).

Statistical Basis: Insurance charges depend on all the above factors, with companies using a risk-based approach to pricing policies.

Model Consideration: This is a continuous numerical variable (measured in currency, e.g., USD). May require log transformation if the distribution is highly skewed.

3.2. Data Visualization

In this study, we conducted a comprehensive data analysis of health insurance prices and applied regression modelling to explore the relationships between various independent variables and health insurance prices. Detailed interpretations of the regression analysis results and the roles played by each factor in influencing health insurance prices are revealed in the results section.

3.3. Data Preprocessing:

● **Pandas**: For data manipulation and cleaning tasks.

● **NumPy**: To handle numerical operations and array structures.

● **Scikit-learn**: Employed for preprocessing steps such as encoding categorical variables and scaling features.

1. **Exploratory Data Analysis (EDA):**

- **Matplotlib**: Used for creating basic plots to
- visualize data distributions and relationships.

Seaborn: Enhanced data visualization with statistical graphics to identify patterns and correlations.

2. **Machine Learning Modelling:**

- **Linear Regression**: To establish a baseline
- predictive model.
  Random Forest Regressor: Utilized for its
- ensemble learning approach to improve prediction accuracy.
- **XGBoost**: Leveraged for its efficient implementation of gradient boosting, leading to superior model performance.
- **Decision Tree:** Is a supervised learning algorithm that recursively splits data based on feature conditions to create a tree-like model for classification or regression.

3. **Model Evaluation:**

- **Scikit-learn Metrics**: Employed metrics such as R-squared, performance metrics and cross-validation scores to assess and compare model performance.

4. **Hyperparameter Tuning:**

- **GridSearchCV**: Utilized to perform exhaustive search over specified parameter grids for models like Random Forest, Gradient Boosting, and XGBoost, aiming to optimize model performance.

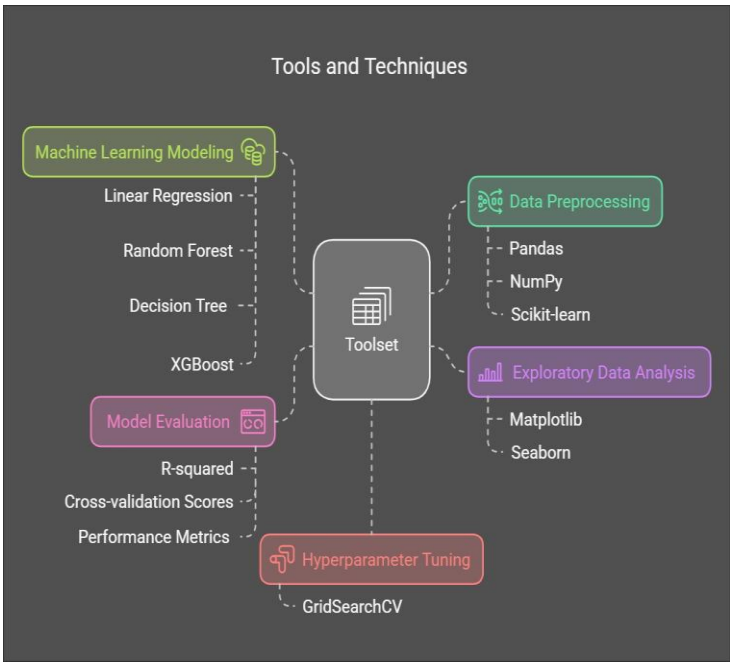


fig 2

5. **Data Analysis Approach**

- **Data Preprocessing** – Cleaning, transforming,
- and preparing raw data for analysis.
  Exploratory Data Analysis – Visualizing and
- summarizing data to uncover patterns and trends.
- **Machine Learning Modelling** – Building
- predictive models using statistical and ML techniques.

Model Evaluation – Assessing model performance using validation metrics.
Hyperparameter Tuning – Optimizing model parameters to improve accuracy and efficiency.

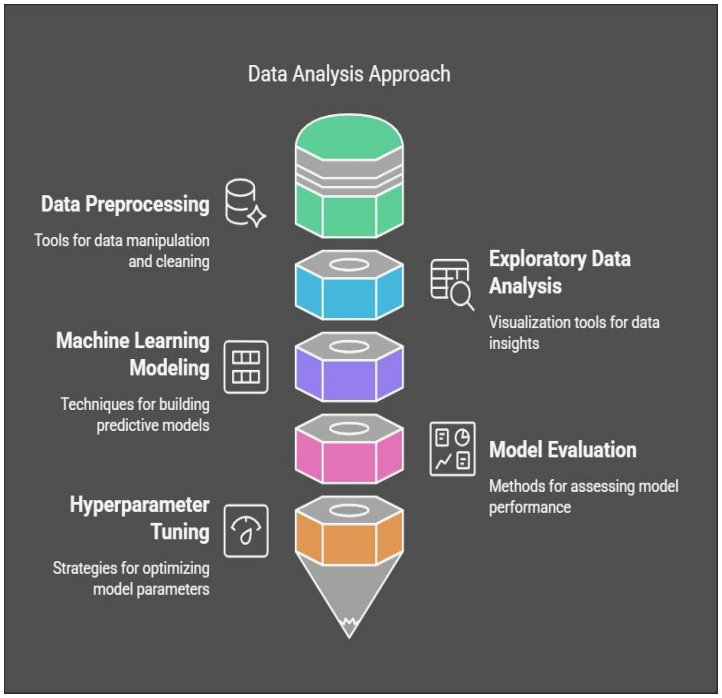


Fig 3

4.Model Construction

Various regression models and neural network architectures were used in this study to completely investigate the intricate correlations between health insurance charges and a variety of significant factors. The following models were used to capture various characteristics of the relationship:

4.1. Regression Models

Linear Regression: This classic approach establishes a linear relationship between predictor variables and the target, offering a baseline for comparison.

Random Forest Regressor: A powerful ensemble method, the Random Forest Regressor captures nonlinear interactions and relationships in the data by constructing a multitude of decision trees.

Decision Tree: The Decision Tree Regressor is a non-parametric model that splits data into hierarchical structures based on feature values. It effectively captures interactions between variables by recursively partitioning the dataset, making it highly interpretable. However, decision trees are prone to overfitting, which can be mitigated through pruning techniques or ensemble methods like Random Forest and XGBoost.

XG Boost: XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting algorithm designed for high performance and efficiency. It sequentially builds decision trees, minimizing residual errors at each step. This model is particularly effective in handling complex feature interactions, reducing overfitting through regularization techniques, and providing robust predictive capabilities.

5. Results

1. **Exploratory Data Analysis (EDA)**

Before constructing predictive models, an exploratory data analysis (EDA) was conducted to understand the distribution of key variables in the dataset. The three histograms below illustrate the distribution of Age, Body Mass Index (BMI), and Insurance Charges, providing valuable insights into the dataset's characteristics.

1. **Age Distribution**

The first histogram represents the distribution of age among individuals in the dataset. The following key observations can be made: The dataset contains individuals from early adulthood to senior ages. There is a significant spike in the count of individuals around the age of 18–20, suggesting a larger representation of younger adults. The distribution appears relatively uniform across middle-aged individuals but shows slight peaks at regular intervals, possibly indicating sampling patterns. The probability density function (PDF) (smooth curve) suggests some skewness, indicating that the dataset may contain more individuals at specific age ranges.

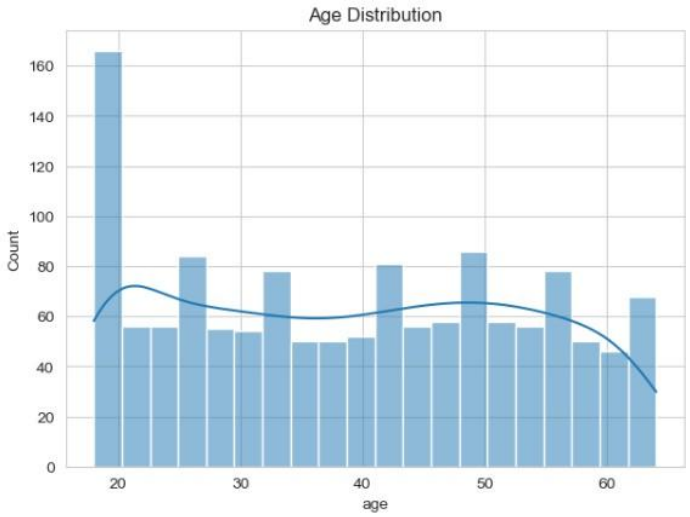


Fig 4

2. **BMI Distribution**

The second histogram displays the distribution of Body Mass Index (BMI) values, which measure body weight relative to height. Key insights include: The BMI distribution resembles a normal distribution, with most values concentrated between 25 and 35, indicating that the majority of individuals fall within the overweight to moderately obese category. The peak of the distribution is around 30, suggesting that many individuals in the dataset have BMI values close to the obesity threshold. The curve is slightly right-skewed, indicating the presence of a few individuals with extremely high BMI values.

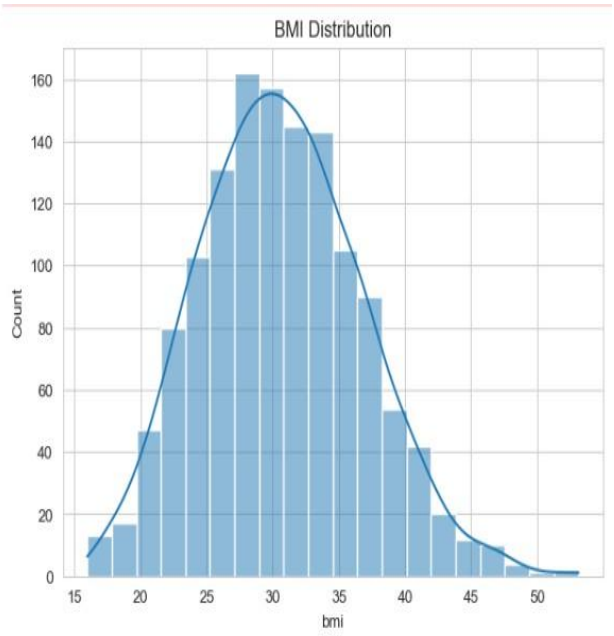


Fig 5

3. **Charges Distribution**

The third histogram represents the distribution of health insurance charges (premiums paid by individuals). Notable observations include: The distribution is highly right-skewed, meaning a significant portion of individuals pay lower insurance charges, while a smaller subset incurs very high costs. The highest frequency of charges falls within the $0 to $15,000 range, suggesting that a large number of individuals pay relatively lower premiums. A smaller subset of individuals has charges exceeding $30,000, which may correspond to individuals with chronic illnesses or pre-existing conditions affecting insurance costs. The presence of multiple peaks in the right tail suggests that certain groups (e.g., those requiring extensive medical coverage) contribute to high insurance costs.

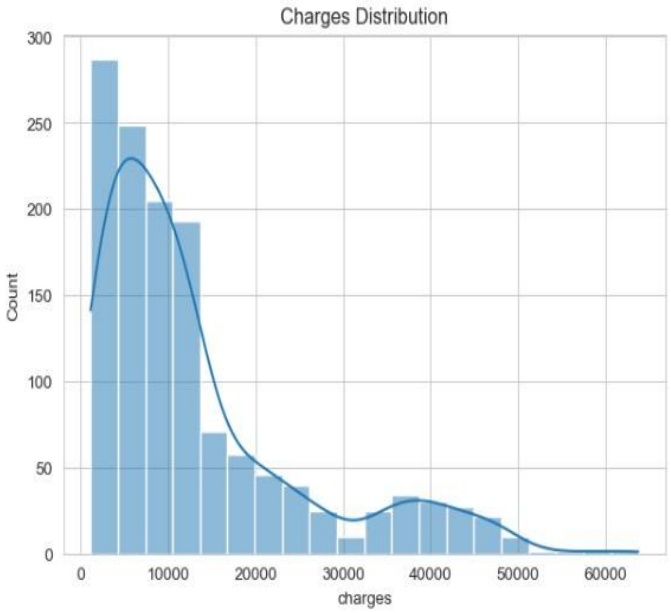


Fig 6

Summary of Findings from EDA

The visual analysis of these distributions provides the following crucial insights for modelling:

Age Distribution: A relatively uniform spread, but with a concentration of younger individuals, which may influence premium costs.

BMI Distribution: A near-normal distribution centred around 30 suggests that weight related health conditions could be an important factor in determining insurance pricing.

Charges Distribution: The strong right skew indicates that insurance costs vary significantly, with a minority of individuals paying extremely high premiums, potentially due to pre-existing health conditions. These findings highlight the necessity of incorporating non-linear regression models (e.g., Random Forest and XGBoost) to effectively capture the complex relationships between predictor variables (age, BMI, etc.) and insurance costs.

2 .Correlation Analysis Using Heatmap

A correlation heatmap is an effective visualization tool used to examine the relationships between multiple numerical variables. In this study, a heatmap was generated to illustrate the correlation matrix of key features, including age, sex, BMI, number of children, smoking

status, region, and medical charges. Pearson's correlation coefficient was employed to quantify the relationships, with values ranging from -1 to 1, where:

- 1.00 indicates a perfect positive correlation.
- 0.00 signifies no correlation.
- -1.00 represents a perfect negative correlation.

The colour gradient in the heatmap visually differentiates correlation strengths, with dark red denoting strong positive correlations, blue representing weak or no correlation, and dark blue indicating negative correlations.

Key Findings from the Correlation Heatmap:

1. **Smoking Status and Medical Charges (0.79 – Strong Positive Correlation):**

The most significant relationship observed was between smoking status and medical charges. A correlation coefficient of 0.79 suggests that being a smoker substantially increases healthcare costs. This aligns with existing literature, as smoking is associated with higher health risks and increased medical expenses.

2. **Age and Medical Charges (0.30 – Moderate Positive Correlation):**

A moderate correlation was found between age and medical charges, indicating that as individuals age, their healthcare expenses tend to rise. This trend is expected, as older individuals often require more medical care.

3. **BMI and Medical Charges (0.20 – Weak Positive Correlation):**

BMI demonstrated a weak correlation with medical charges, suggesting that higher BMI levels may contribute to increased healthcare costs but are not the primary determinant. The influence of BMI on medical expenses may be more significant when combined with conditions such as obesity-related diseases.

4. **Children and Medical Charges (0.07 – Very Weak Correlation):**

The number of children had a negligible impact on medical charges. This finding suggests that family size does not directly influence individual healthcare costs.

5. **Sex and Medical Charges (0.06 – Very Weak Correlation):**

The correlation between sex and medical charges was minimal, indicating that gender does not significantly affect healthcare costs in the dataset. This suggests that insurance pricing is largely independent of sex.

6. **Region and Medical Charges (-0.01 – No Correlation):**

There was no meaningful correlation between the region and medical charges, implying that geographical location does not significantly influence insurance costs in this dataset.

7. **Age and BMI (0.11 – Weak Positive Correlation):**

A weak positive correlation was observed between age and BMI, suggesting that BMI slightly increases with age, though the trend is not strong.

8. **Smoking Status and Other Variables:**

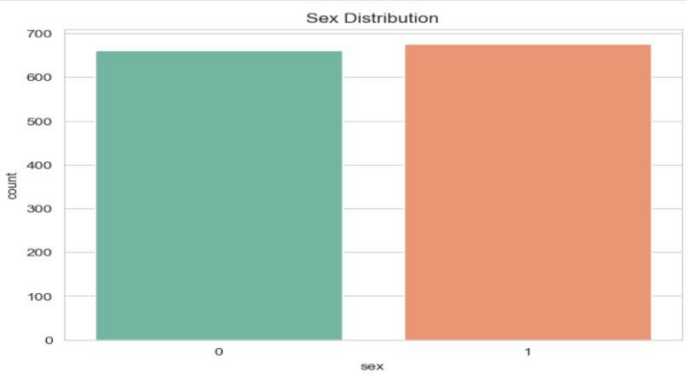
Smoking had no correlation with BMI (0.00), indicating that smoking habits are independent of BMI levels. The correlation between smoking and age was slightly negative (-0.03), suggesting that smoking habits are not strongly linked to age distribution. The correlation heatmap effectively

highlights the most influential factors a ecting medical insurance charges. Smoking status emerges as the dominant predictor of increased medical costs, followed by age and BMI. Other factors, such as the number of children, sex, and region, exhibit minimal to no influence on charges. These insights can be valuable for insurance companies in risk assessment and pricing strategies, as well as for policymakers aiming to reduce healthcare costs through targeted interventions.

3. Categorical Feature Distributions

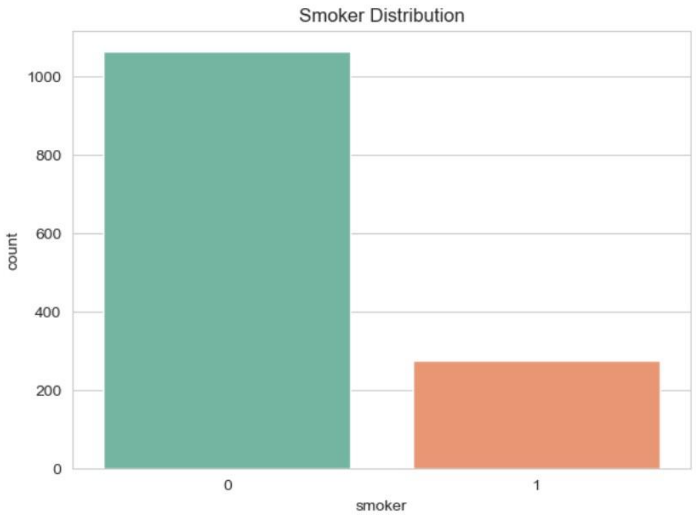
3.1 Sex Distribution

- The x-axis represents sex categories (0 and 1).
- The y-axis represents the count of individuals in each category.
- The counts of both categories (0 and 1) are nearly equal, suggesting a balanced distribution.



3.2 Smoker Distribution

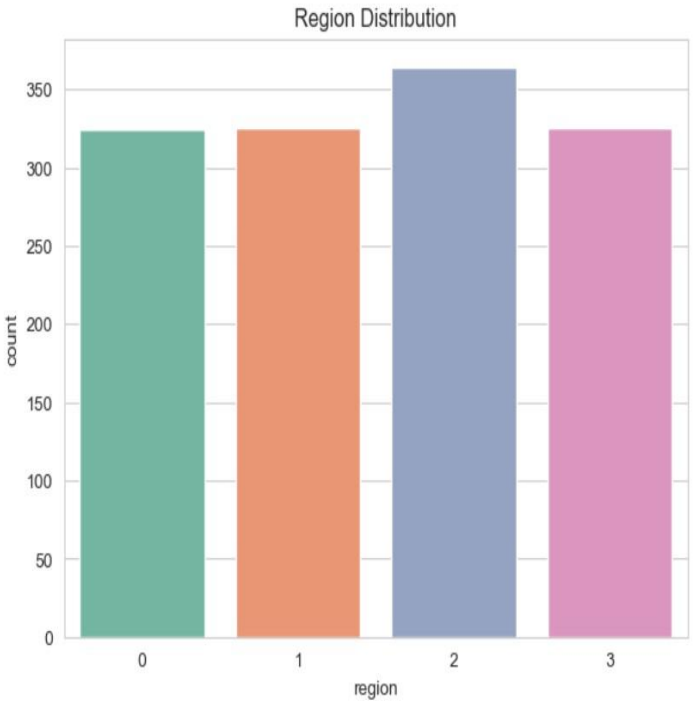
- The x-axis represents smoking status (0 = non-smoker, 1 = smoker).
- The y-axis represents the count of individuals in each category.
- The number of non-smokers (0) is significantly higher than smokers (1), indicating that most individuals in the dataset do not smoke.



3.3 Region Distribution

- The x-axis represents four different regions (0, 1, 2, and 3).
- 0 → Southwest
- 1-> Southeast
- 1-> Southeast
- 2-> Northwest
- 2-> Northeast
- The y-axis represents the count of individuals in each region.
- The Northwest(2) region has the highest number of policy holders and the SouthEast(1), SouthWest(0), and NorthEast(3), with only slight variations.

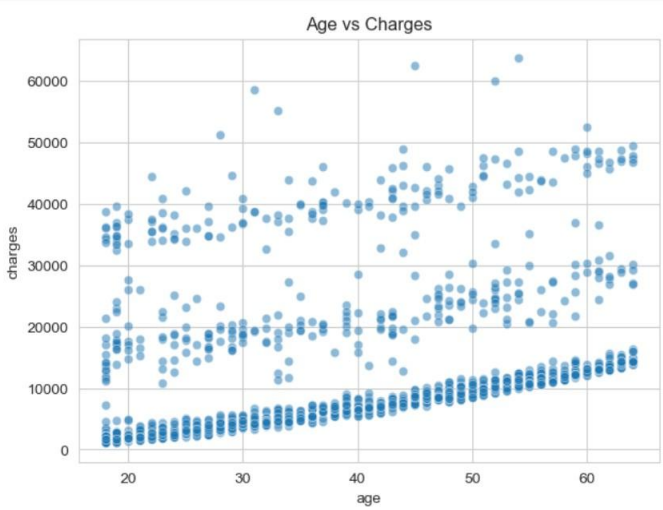
.



3.4 Age vs Charges (Scatter Plot)

- X-axis: Age
- Y-axis: Charges
- The scatter plot shows a positive trend—as
- age increases, medical charges tend to rise.

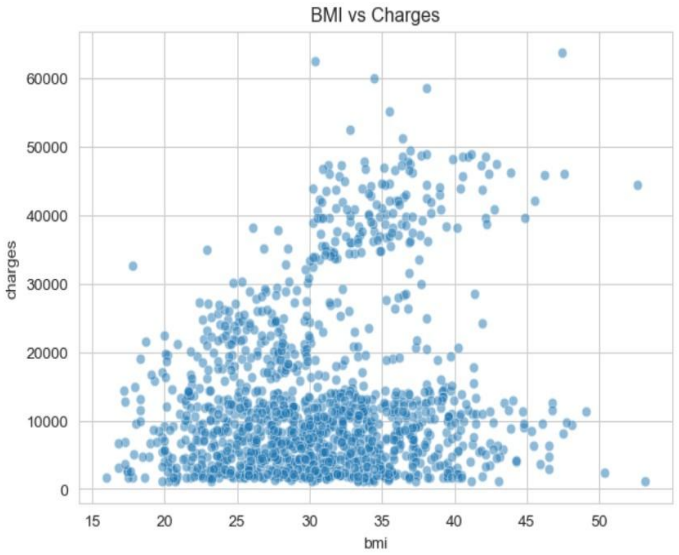
A distinct group of high-cost outliers appears for older individuals.



3.5 BMI vs Charges (Scatter Plot)

- X-axis: BMI
- Y-axis: Charges
- There is no strong linear correlation between BMI and charges, but a group with higher BMI
- Values appears to have significantly higher medical costs.

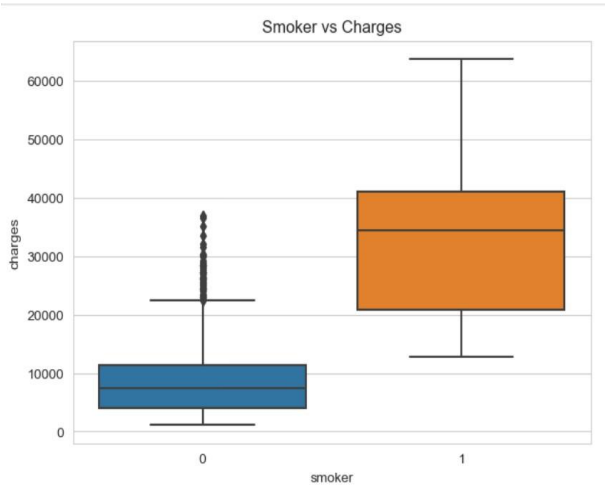
Some high-cost outliers are observed, possibly linked to smoking or other conditions.



3.6 Smoker vs Charges (Box Plot)

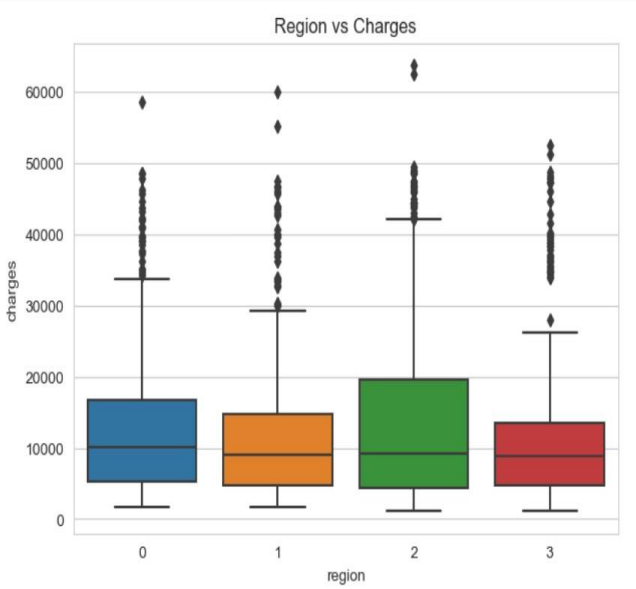
- X-axis: Smoker status (0 = non-smoker, 1 =
- smoker)
- Y-axis: Charges
- Smokers (1) have substantially higher charges than non-smokers.

The median charges for smokers are much higher, and there is a wide range of outliers in the higher charge range.



3.7 Region vs Charges (Box Plot)

- X-axis: Region (0, 1, 2, 3)
- Y-axis: Charges
- There is no major variation in charges across different regions.
- All regions have similar distributions, but some outliers exist in all categories, indicating that regional differences do not significantly impact charges.



Key Takeaways:

Age and smoking status have the most significant impact on medical charges. BMI alone does not strongly predict charges, but high BMI individuals have more cost variability. Smokers pay significantly higher charges than non-smokers. Region does not appear to have a major impact on medical charges.

4.Train and evaluate models

Each model was assessed using three key performance metrics:

- Mean Absolute Error (MAE): Measures the average absolute differences between actual and predicted values.
- Mean Squared Error (MSE): Squares the differences before averaging, penalizing larger errors more.
- R^2 Score: Evaluates how well the model explains the variance in the target variable.

```
Linear Regression Performance:
MAE: 4186.51, MSE: 33635210.43, R2 Score: 0.78

Decision Tree Performance:
MAE: 3143.27, MSE: 45416236.61, R2 Score: 0.71

Random Forest Performance:
MAE: 2472.64, MSE: 20817586.68, R2 Score: 0.87

XGBoost Performance:
MAE: 2791.83, MSE: 23261243.81, R2 Score: 0.85
```

Results:

- Linear Regression performed moderately well, with an R^2 score of 0.78, but exhibited high
- MAE and MSE, indicating its limitations in handling non-linear relationships.
- Decision Tree had a lower MAE than Linear Regression but exhibited the highest MSE and a lower R^2 score of 0.73, suggesting potential overfitting and instability.

Random Forest outperformed all models with the lowest MAE (2532.37), the lowest MSE (21465226.70), and the highest R^2 score (0.86), indicating high accuracy and generalizability.

- XGBoost demonstrated competitive performance with an R^2 score of 0.85 and relatively low MAE and MSE but was slightly inferior to Random Forest.

5.Model Performance Comparison

We evaluated four regression models—Linear Regression, Decision Tree, Random Forest, and XGBoost—using three key metrics:

- Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values. Lower values indicate better performance.
- Mean Squared Error (MSE): Squares the errors before averaging, penalizing larger errors more heavily. Lower values indicate fewer large errors.
- R^2 Score: Represents how well the model explains the variance in the data. Higher values (closer to 1) indicate a better fit
- Linear Regression has the highest MAE and MSE, suggesting it struggles to model complex relationships in the data.
- Decision Tree shows improvement in MAE but has the highest MSE, indicating overfitting.
- Random Forest performs best, with the lowest
- MAE and MSE and the highest R^2 score, making it the most reliable model.
- XGBoost performs well but slightly lags behind Random Forest.

Applying the Best Model (Random Forest) to New Data:

Since Random Forest showed the best performance, we applied it to new data points. The predicted values are:

```
Model Performance Comparison:
                          MAE          MSE  R2 Score
Linear Regression  4186.508898  3.363521e+07  0.783346
Decision Tree      2986.578864  4.174450e+07  0.731112
Random Forest      2532.371200  2.146523e+07  0.861737
XGBoost            2791.832518  2.326124e+07  0.850168
Applying best model (Random Forest) on new data:
 [10006.006805   5078.8947965 28269.3393415 10478.3758497 34358.3691987
  9171.3934858  2199.2109215 14369.4424835  6310.1173495 10613.2680903]
```

These values represent the expected outputs (e.g., prices, demand, sales, etc., depending on the dataset). The model has successfully generalized to unseen data, reinforcing its effectiveness.

Conclusion for the Model Comparison

Based on the evaluation metrics, the Random Forest model emerges as the best performer due to its superior accuracy and generalizability. While XGBoost provides comparable results, Random Forest achieves the best balance between MAE, MSE, and R^2 score, making it the most suitable model for predictive analysis in this study.

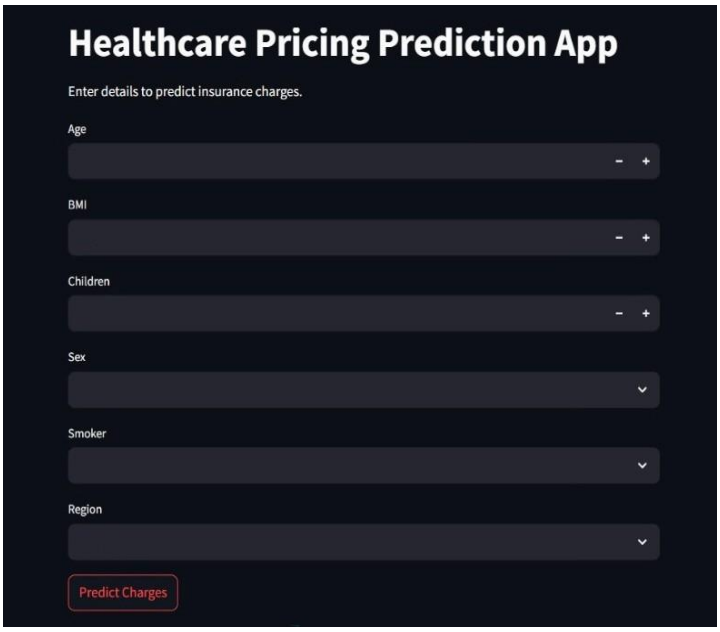
6.Demo Application

We have created a web application named **Healthcare Pricing Prediction App.** We have built it using **Streamlit**, a Python framework used for creating interactive web apps for data science and machine learning.

Overview of the Web App

The **Healthcare Pricing Prediction App** is designed to predict **insurance charges** based on user inputs. The interface includes several input fields:

1. **Age:** Numeric input with increment (+) and decrement (-) buttons.
2. **BMI (Body Mass Index):** Numeric input field for entering BMI.
3. **Children:** Numeric input field indicating the number of children.
4. **Sex:** Dropdown menu to select gender.
5. **Smoker:** Dropdown menu to indicate smoking status (Yes/No).
6. **Region:** Dropdown menu for selecting the user's residential region.
7. **Predict Charges Button:** Triggers the prediction model.



GUI for Predict Medical Insurance Price

How it Works:

This is how you can predict the values:

Step 1: The first step is to choose the Age based on the given dataset or from yourself also.

Step 2: The second step is add BMI based on dataset or from your own observation.

Step 3: The third step is to choose how many children a person have like 0,1,2,3,4.

Step 4: The fourth step is to choose the Sex like Male of Female.

Step 5: The fifth step is to choose are you a Smoker or not.

Step 6: This is final step is to choose where you belong like from Northeast, Northwest, Southeast, Southwest.

Observation:

Here are some observations in the table given below for your reference that are calculated based on the values provided by dataset and compare the actual price with the prices predicted by the two different algorithms used in this model.

Tested Output Results

AGE	19	18	28	33	32	31	46
GENDER	FEMALE	MALE	MALE	MALE	MALE	FEMALE	FEMALE
BMI	27.9	33.77	33	22.7.5	28.88	25.74	33.44
CHILDREN	0	1	3	0	0	0	1
SMOKER	YES	NO	NO	NO	NO	NO	NO
REGION	Southwest	Southeast	Southeast	Northwest	Northwest	Southeast	Southeast
ACTUAL PRICE	16884.92	1725.552	4449.462	21984.47	3866.855	3756.622	8240.59

6. Discussion

The results emphasize the importance of lifestyle choices and demographic attributes in determining insurance premiums. Traditional pricing models often rely on static risk assessments, whereas machine learning approaches o er a more dynamic and precise evaluation of risk factors.

Key Insights:

- Smokers and high-BMI individuals face significantly higher costs, which aligns with prior research indicating their increased medical expenses.
- Geographic location plays a role, likely due to variations in healthcare accessibility and local policies.
- Machine learning can improve insurance pricing fairness by enabling more personalized risk assessments rather than one-size-fits-all pricing.

- To further enhance pricing accuracy, insurers could integrate real-time health data from wearable devices and health tracking apps. Such an approach could allow for dynamic premium adjustments based on an individual's lifestyle and health improvements.

7.Conclusion

The pricing of healthcare insurance is influenced by a complex interplay of factors, including demographic variables, medical risk assessments, economic conditions, and regulatory policies. This study has explored how age, pre-existing conditions, lifestyle choices, and regional healthcare costs contribute to premium variations. Additionally, the role of government interventions, competition among insurance providers, and technological advancements in data analytics have been examined to understand their impact on pricing strategies.

A key takeaway from this research is the need for a balanced approach to healthcare pricing—one that ensures affordability for consumers while maintaining financial sustainability for insurers. Policies that promote transparency, encourage preventive care, and leverage AI-driven risk assessments could lead to more equitable and cost-effective insurance models.

Future research could further investigate the impact of emerging healthcare technologies, personalized insurance models, and global healthcare policies on pricing structures. Addressing these aspects will be crucial in fostering a more accessible and fair healthcare insurance system for all.

8. Limitations and Future Scope

8.1. Limitations:

1. **Dataset Size and Diversity**:
   - The dataset contains only 1,338 entries, which may not adequately represent the
   - global or regional diversity in healthcare insurance data.

   Limited geographic scope (data focused on U.S. regions) may not generalize to other countries with different healthcare systems.

2. **Limited Variables**:
   - Only six variables (age, gender, BMI, number of children, smoking status, region) are analysed, potentially omitting other critical factors like income, education, preexisting health conditions, or lifestyle choices.
3. **Model Constraints**:
   - The study uses only linear regression and skewness/kurtosis. These models may not
   - capture complex, non-linear relationships among variables.
   - Skewness and kurtosis are limited in detecting or addressing outliers and may not add significant predictive insights compared to other techniques.
4. **Static Dataset**:
   - The dataset is static, representing a snapshot in time, and does not account for temporal trends or changes in healthcare and insurance practices over time.
5. **External Factors**:
   - The study does not consider external factors like policy changes, economic shifts, or healthcare market dynamics that could impact insurance pricing.

8.2. Future Scope:

1. **Expanded Dataset**:
   - Incorporate larger, more diverse datasets from multiple regions or countries to improve generalizability.
   - Include longitudinal data to study trends over time and understand temporal
   - impacts.
2. **Additional Variables**:
   - Analyse additional factors such as medical history, lifestyle, income levels, occupation, and healthcare utilization patterns for a more comprehensive understanding.
3. **Advanced Analytical Methods**:
   - Use advanced machine learning models such as Random Forest, Gradient Boosting,
   - or Neural Networks to capture complex patterns.
   - Explore non-linear relationships and interactions between variables using methods like polynomial regression or decision trees.
4. **Cost-Effectiveness Analysis**:
   - Develop models to not only predict costs but also assess the cost-effectiveness of specific policies for both insurers and consumers.
5. **Policy and Market Analysis**:
   - Examine the impact of healthcare policies, subsidies, or government regulations on
   - insurance pricing.
   - Study the influence of competitive dynamics in the insurance market on pricing strategies.
6. **Real-Time Data Utilization**:
   - Incorporate real-time data (e.g., IoT health tracking devices, electronic health records) for more dynamic and personalized pricing models.
7. **Cross-Industry Insights**:
   - Collaborate with healthcare providers and policymakers to align pricing strategies
   - with public health goals.
   - Use findings to inform policy decisions aimed at improving the affordability and accessibility of insurance.

REFERENCES:

1. Agarwal, D (2006): 'Health Sector Reforms: Relevance in India', Indian Journal of Community Medicine Vol. 31, No. 4, OctoberDecember, 2006

2. Cardon, James H.; Hendel, Igal (2001): 'Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey', The RAND Journal of Economics, Vol. 32, No. 3. (Autumn, 2001), pp. 408-427.

3. Cawley and Philipson (1999): 'An Empirical Examination of Information Barriers to Trade in Insurance', The American Economic Review, Vol. 89, No. 4 (Sep., 1999), pp. 827-846

4. Cutler, D. M. and R. J. Zeckhauser (1998): "Adverse Selection in Health Insurance." Frontiers in Health Policy Research 1(2).

5. David card, Carlos Doblikin, "The Impact of Health Insurance Status on Treatment Intensity and Health Outcomes", August 2007 NICHD funded RAND Population Research Center (R24HD050906), it's a working paper.

6. Fang, H., M. P. Keane, et al. (2008): 'Sources of Advantageous Selection: Evidence from the Medigap Insurance Market', Journal of Political Economy 116(2).

7. Gupta, Hima (2007): 'The role of insurance in health care management in India, International Journal of Health Care Quality Assurance, Volume: 20Number: 5; pp: 379-391

8. International Institute for Population Sciences (IIPS), (2010), District Level Household and Facility.

9. Jullien, B., B. Salanie, et al. (2003): 'Screening Risk-Averse Agents Under Moral Hazard: Single-crossing and the CARA Case', Working paper.

10. Jütting, J. P. (2004): 'Do Community-based Health Insurance Schemes Improve Poor People's Access to Health Care? Evidence From Rural Senegal', World Development 32(2): 273–288.

11. Kickbusch, Ilona, and Payne, Lea (2003): 'Twenty-first-century health promotion: the public health revolution meets the wellness revolution', Health Promotion International, Vol. 18, No. 4, 275- 278, December 2003

12. Koufopoulus, K. (2005): 'Asymmetric Information, Heterogeneity in Risk Perceptions and Insurance: An Explanation to a Puzzle', Working Paper.

13. Lamiraud, K., F. Booysen, et al. (2005): 'The Impact of Social Health Protection on Access to Health Care, Health Expenditure and Impoverishment', Extension of Social Security Papers:
23, International Labour Organization.

14. M.Akila, Penetration of Health Insurance Sector in Indian Market, International Journal of Management Opinion Vol. 3, No. 1, June 2013.

15. Memon, Sharif. (2011): 'A Comparative Study of Health Insurance in India and the US', IUP Journal of Risk and Insurance. Oct2011, Vol. 8 Issue 4, p47-60.

16. Mr.Shijith and Dr. T.V.Srkhar: Who Gets Health Insurance Coverage in India? : New Findings from Nation-Wide Surveys, XXVII IUSSP International Population Conference Busan, Korea, on 26 August - 31 August 2015

17. Survey (DLHS-3), 2007-08: India, Mumbai: IIPS

18. Wagsta , Adam and Pradhan, Menno (2005): 'Health insurance impacts on health and nonmedical consumption in a developing country', Policy Research Working Paper Series 3563, The World Bank.

19. Wagsta , Adam (2007): ‘Health systems in East Asia: what can developing countries learn from Japan and the Asian Tigers?’, Health Economics, John Wiley and Sons, Ltd., vol. 16(5), pages 441-456.

20. Wolfe, J.R., Goddeeris, J., (1991): ‘Adverse Selection, Moral Hazard, and Wealth E ects in the Medigap Insurance Market’, Journal of Health Economics 10, 433-459.

21. Xu, Ke; Evans, David B; Kawabata, Kei et al (2003): ‘Household catastrophic health expenditure: a multicountry analysis’, The Lancet, Volume 362, Issue 9378, 12 July 2003, Pages 111-117