

**SCHOOL OF COMPUTER SCIENCE**

**CASE STUDY (Weightage 30%)**

**JAN 2025 SEMESTER**

<b>MODULE NAME</b>	<b>: Statistical Inference and Modelling</b>
<b>MODULE CODE</b>	<b>: ITS66804</b>
<b>DUE DATE</b>	<b>: Week 11</b>
<b>PLATFORM</b>	<b>: MyTIMES</b>

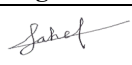

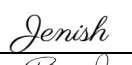
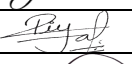
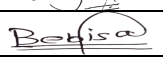
**This paper consists of TEN (10) pages, inclusive of this page.**

**Group No:**

**Project Title:**

***STUDENT DECLARATION***

- I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignment and of the guidance issued by the School of Computing and IT concerning plagiarism and proper academic practice, and that the assessed work now submitted is in accordance with this regulation and guidance.*
- I understand that, unless already agreed with the School of Computing and IT, assessed work may not be submitted that has previously been submitted, either in whole or in part, at this or any other institution.*
- I recognise that should evidence emerge that my work fails to comply with either of the above declarations, then I may be liable to proceedings under Regulation*

<b>No</b>	<b>Student Name</b>	<b>Student ID</b>	<b>Date</b>	<b>Signature</b>	<b>Score</b>
1.	Sahel Shrestha	0370021	18 <sup>th</sup> March		
2.	Sujal G.C	0370039	18 <sup>th</sup> March		
3.	Jenish Karmacharya	0370033	18 <sup>th</sup> March		
4.	Riya Maharjan	0369932	18 <sup>th</sup> March		
5.	Bebisa Regmi	0370002	18 <sup>th</sup> March		

## Marking Rubric

Group Assignment Marking Rubrics					
<b>Abstract (5 marks)</b>	5 marks A clear and concise abstract that gives the reader a clear idea of what the project is about and why it is interesting. The following components need to be included  i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	4 marks A clear abstract that gives the reader a clear idea of what the project is about. Four of the following components are included  i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	The abstract is difficult to read and/or is very vague and/or doesn't sell the project as well as it might have. Three of the following components are included  i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	2 marks Unable to read the abstract and/or is very vague and/or doesn't sell the project as well as it might have. Only two of the following components are included  i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion	1 mark Unable to read the abstract. Only one of the following components is included  i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion

<b>Introduction (10 marks)</b>	9-10 marks A readable write-up that explains what the problem is and why it is of interest. The following components need to be	7-8 marks A readable write-up that explains what the problem is. Three of the following components are included. i. Problem ii. Negative	5-6 marks The write-up is difficult to read, somewhat vague, or doesn't make a really good case for why the problem is of interest.	3-4 marks Unable to read the write-up and/or is very vague. Only one of the following components are included. i. Problem	1-2 marks Unable to read the write-up. None of the following components are included. i. Problem ii. Negative impact of the
------------------------------------	---	--	---	---	---

	included i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem	impact of the problem iii. Parties affected iv. Benefit of solving the problem	Two of the following components are included. i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem	ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem	problem iii. Parties affected iv. Benefit of solving the
--	---	--	--	---	--

<b>Literature Review (20marks)</b>	18-20 marks An outstanding overview, with an insightful analysis of prior work and a clear connection between prior work and the proposed method. The following components are given. (8 articles) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	15- 17 marks A comprehensive overview of prior work that gives the reader a clear idea of what's out there and how the proposed method is different. Four of the following components are given. (6 articles) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	10-14 marks A fairly good overview of prior work, and some connection is made to the proposed method. Three of the following components are given. (5 articles) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	5-9 marks An overview of several papers related to the proposed method, and some attempt is made to connect the prior work to the current method. Two of the following components are given. (4 marks) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	1-4 marks An overview of several related papers, but not within a coherent conceptual frame-work. One of the following components are given. (2 marks) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review
<b>Data (5marks)</b>	5 marks The data are comprehensive	4 marks The data are fairly	3 marks The data are not	2 marks The explanations	1 mark The explanations

	<p>and clearly described. At least 6 of the following components are given.</p> <p>i. Source of the data</p> <p>ii. Description of the data and its context</p> <p>iii. Statistics of the data</p> <p>iv. Presentation, visualization and quantification of the data and images</p> <p>v. Conclusion</p>	<p>explained. At least 5 of the following components are given.</p> <p>i. Source of the data</p> <p>ii. Description of the data and its context</p> <p>iii. Statistics of the data</p> <p>iv. Presentation, visualization and quantification of the data and images</p> <p>v. Conclusion</p>	<p>comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given.</p> <p>i. Source of the data</p> <p>ii. Description of the data and its context</p> <p>iii. Statistics of the data</p> <p>iv. Presentation, visualization and quantification of the data and images</p> <p>v. Conclusion</p>	<p>are significantly flawed. At least 3 of the following components are given.</p> <p>i. Source of the data</p> <p>ii. Description of the data and its context</p> <p>iii. Statistics of the data</p> <p>iv. Presentation, visualization and quantification of the data and images</p> <p>v. Conclusion</p>	<p>are flawed. At least 2 of the following components are given.</p> <p>i. Source of the data</p> <p>ii. Description of the data and its context</p> <p>iii. Statistics of the data</p> <p>iv. Presentation, visualization and quantification of the data and images</p> <p>v. Conclusion</p>
--	--	--	--	---	---

<b>Metho d (20 marks)</b>	17-20 marks The methods of analysis Are comprehensive and clearly described. At least 6 of the following components are given.  i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research	13-16 marks The methods of analysis are fairly explained. At least 5 of the following components are given.  i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research objective v. Information	9-12 marks The methods of analysis are not comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given.  i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods	5-8 marks The methods of analysis are significantly flawed. At least 3 of the following components are given.  i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods	1-4 marks The methods of analysis are flawed. At least 2 of the following components are given.  i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis iv. Statistical methods address the research objective
-----------------------------------	--	--	--	--	---

	objective v. Information on data analysis process vi. Clear relationship between methods	on data analysis process vi. Clear relationship between methods	address the research objective v. Information on data analysis process vi. Clear relationship between methods	address the research objective v. Information on data analysis process vi. Clear relationship between methods	v. Information on data analysis process vi. Clear relationship between methods
--	---	---	--	---	--

<b>Result &amp; Discussion (20 marks)</b>	<p>17-20 marks</p> <p>The results are comprehensive and clearly described. At least 6 of the following components are given.</p> <p>i. Subheadings are included and are clear and informative</p> <p>ii. Figures and tables are supported by text</p> <p>iii. Correct interpretation of the results</p> <p>iv. Results with tables and diagrams</p> <p>v. Additional insight to the content</p> <p>vi. Critical analysis of the results</p> <p>vii. Clearly addresses the research question</p>	<p>13-16 marks</p> <p>The results are fairly explained. At least 5 of the following components are given.</p> <p>i. Subheadings are included and are clear and informative</p> <p>ii. Figures and tables are supported by text</p> <p>iii. Correct interpretation of the results</p> <p>iv. Results with tables and diagrams</p> <p>v. Additional insight to the content</p> <p>vi. Critical analysis of the results</p> <p>vii. Clearly addresses the research question</p>	<p>9-12 marks</p> <p>The results are not comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given.</p> <p>i. Subheadings are included and are clear and informative</p> <p>ii. Figures and tables are supported by text</p> <p>iii. Correct interpretation of the results</p> <p>iv. Results with tables and diagrams</p> <p>v. Additional insight to the content</p> <p>vi. Critical analysis of the results</p> <p>vii. Clearly addresses the</p>	<p>5-8 marks</p> <p>The results are significantly flawed. At least 3 of the following components are given.</p> <p>i. Subheadings are included and are clear and informative</p> <p>ii. Figures and tables are supported by text</p> <p>iii. Correct interpretation of the results</p> <p>iv. Results with tables and diagrams</p> <p>v. Additional insight to the content</p> <p>vi. Critical analysis of the results</p> <p>vii. Clearly addresses the research question</p>	<p>1-4 marks</p> <p>The results are flawed. At least 2 of the following components are given.</p> <p>i. Subheadings are included and are clear and informative</p> <p>ii. Figures and tables are supported by text</p> <p>iii. Correct interpretation of the results</p> <p>iv. Results with tables and diagrams</p> <p>v. Additional insight to the content</p> <p>vi. Critical analysis of the results</p> <p>vii. Clearly addresses the research question</p>
---	---	--	--	--	--

			research question		
--	--	--	-------------------	--	--

<b>Limitation and future Study (10 marks)</b>	<p>9-10 marks An insightful and correct analysis. The following components are given.</p> <p>i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies</p>	<p>7-8 marks A correct analysis that could be more complete and is not very insightful. One of the following components is missing.</p> <p>i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies</p>	<p>5-6 marks An incomplete or somewhat incorrect analysis. Two of the following components are missing.</p> <p>i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies</p>	<p>3-4 marks An incorrect analysis. One of the following components are given.</p> <p>i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies</p>	<p>1-2 marks No analysis. None of the following components are given.</p> <p>i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies</p>
---	--	--	--	---	--



<b>Conclusion (5 marks)</b>	<p>5 marks A clear and insightful summary of the paper, perhaps with interesting ideas for future work. The following components are given.</p> <p>i. Restate your research topic</p>	<p>4 marks A summary of the experiment is given, but the conclusion is a mere summary. The ideas for future work are not interesting. One of the following components is</p>	<p>3 marks A flawed conclusion. Two of the following components are missing.</p> <p>i. Restate your research topic ii. Restate the objective iii. Summarize the main topics</p>	<p>2 marks An incorrect conclusion. Three of the following components are missing.</p> <p>i. Restate your research topic ii. Restate the objective iii. Summarize</p>	<p>1 marks No conclusion. One of the following components is given.</p> <p>i. Restate your research topic ii. Restate the objective iii. Summarize</p>
	<p>ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts</p>	<p>missing. i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts</p>	<p>iv. Significance of results v. Conclude the thoughts</p>	<p>the main topics iv. Significance of results v. Conclude the thoughts</p>	<p>the main topics iv. Significance of results v. Conclude the thoughts</p>

<b>Format (5 marks)</b>	<p>5 marks A clear and correct formatting. The following components are given.</p> <p>i. Number of pages 10 -15  ii. Use the correct template  iii. Similarity index less than 20%  iv. All the sections given in proper order  v. Readable pdf file</p>	<p>4 marks A clear and correct formatting . One of the following components is missing.</p> <p>i. Number of pages 10 -15  ii. Use the correct template  iii. Similarity index less than 20%  iv. All the sections given in proper order  v. Readable pdf file</p>	<p>3 marks Two of the following components are missing.</p> <p>i. Number of pages 10 -15  ii. Use the correct template  iii. Similarity index less than 20%  iv. All the sections given in proper order  v. Readable pdf file</p>	<p>2 marks Three of the following components are missing.</p> <p>i. Number of pages 10 -15  ii. Use the correct template  iii. Similarity index less than 20%  iv. All the sections given in proper order  v. Readable pdf file</p>	<p>1 marks One of the following components is given.</p> <p>i. Number of pages 10 -15  ii. Use the correct template  iii. Similarity index less than 20%  iv. All the sections given in proper order  v. Readable pdf file</p>
-------------------------	--	---	---	---	--

## Table of Contents

<b>Marking Rubric .....</b>	<b>2</b>
<b>Abstract.....</b>	<b>12</b>
<b>Introduction.....</b>	<b>14</b>
<b>Background .....</b>	<b>14</b>
<b>Problem Statement.....</b>	<b>14</b>
<b>Research Objectives.....</b>	<b>14</b>
Scope of the study .....	15
Conclusion .....	15
<b>LITERATURE REVIEW.....</b>	<b>16</b>
Differentiation of Contribution: .....	17
<b>Data .....</b>	<b>18</b>
<b>Source of data .....</b>	<b>19</b>
<b>Methods.....</b>	<b>20</b>
<b>Exploratory Data Analysis (EDA) .....</b>	<b>20</b>
<b>Statistical Data Analysis Methods .....</b>	<b>21</b>
<b>Data Cleaning and Preprocessing.....</b>	<b>22</b>
<b>Statistical and Machine Learning Methods.....</b>	<b>22</b>
<b>Information on Data Analysis Process .....</b>	<b>23</b>
<b>Relationship Between Methods and Research Objectives .....</b>	<b>23</b>
<b>Results and Discussion.....</b>	<b>24</b>
<b>Exploratory Data Analysis .....</b>	<b>24</b>
Categorical Variables .....	24
Numerical Variables and Insights .....	26
Boxplot of charges by Smoker, Region, BMI category, Gender .....	27
Scatter plot of BMI vs. Charges by Smoker Status .....	29
Correlation heatmap.....	30
<b>Statistical Analysis .....</b>	<b>30</b>
T-tests .....	30
ANOVA.....	32
Linear Regression: .....	33
<b>Predictive Modeling.....</b>	<b>34</b>

Data Splitting .....	34
□ Decision Tree Model.....	34
□ Random Forest Model: .....	35
Confusion matrix .....	35
F1-score, precision, recall .....	36
Accuracy .....	37
<b>Critical Analysis .....</b>	<b>37</b>
<b>Related Questions.....</b>	<b>38</b>
<b>Limitations and Future Study.....</b>	<b>38</b>
<b>Conclusion .....</b>	<b>40</b>
<b>Appendix.....</b>	<b>41</b>
<b>References .....</b>	<b>41</b>

## Figures of Table

Fig 1: Data Exploration.....	21
Fig 2: Distribution of Gender.....	24
Fig 3: Distribution of Smoker variable .....	25
Fig 4: Distribution of Region variable .....	25
Fig 5: Distributions of Numerical variables.....	26
Fig 6: Boxplots of different variables .....	28
Fig 7: Scatter plot of BMI vs. Charges by Smoker Status .....	29
Fig 8: Correlation heatmap .....	30
Fig 9: T-test for charges by smoker.....	31
Fig 10: T-test for charges by sex .....	31
Fig 11: Check BMI significant using ANOVA .....	32
Fig 12: Check Region significant using ANOVA .....	33
Fig 13: Linear Regression.....	33
Fig 14: Data Splitting.....	34
Fig 15: Decision Tree Model .....	34
Fig 16: Confusion Matrix for Decision Tree.....	34
Fig 17: Random Forest Model .....	35
Fig 18: Confusion Matrix for Random Forest .....	35
Fig 19: Visualization of Confusion Matrices .....	35
Fig 20: F1-score, Recall, Precision of both models .....	37
Fig 21: Accuracy of both models .....	37

## Abstract

People and families require healthcare insurance to protect themselves from unexpected medical expenses. Many factors including location and personal choices and population demographics influence how much insurance costs will be (Cevolini & Esposito, 2020). Understanding the price variations helps to create better models and achieve healthcare price equality while driving down healthcare expenses (ul Hassan et al., 2021).

The analysis of healthcare insurance pricing examines a dataset through the examination of age, gender, BMI, dependent counts, smoking habits, geographical and insurance costs variables. A statistical modeling analysis together with exploratory data analysis (EDA) reveals the existence of strong relationships between specific variables and insurance charges.

Results show that smoking status acts as the leading factor which determines insurance premiums. Non-smokers pay insurance costs 4.5 times lower than what smokers need to pay (Hanafy & Mahmoud, 2021). Smoking results in elevated heart danger and respiratory problems and multiple long-term medical complications thus affecting insurance costs. Individuals with higher BMIs need to pay more insurance premiums because they are likely to get diabetes and high blood pressure as well as other obesity-linked disorders (Wichmann & Eberl, 2022).

Insurance premiums increase dramatically as individuals become older so the dependency between age and insurance costs demonstrates a direct and powerful positive relationship. Older age carries higher insurance costs because age itself links to higher risks for chronic diseases and various health problems. Research shows that people under 30 years pay less for premiums yet those above 50 encounter greater medical costs.

Rates assessed by insurance companies depend heavily on the geographic area where customers reside. Residents of the Southeast bear the highest insurance costs while those dwelling in the northwest, northeast and southwest sectors follow. The rates determine their values according to health costs and insurance business price methods in addition to healthcare service capabilities in particular areas.

Numerical factors such as family dependents prove to impact insurance premiums to a lesser degree. Insurer health premiums rise with each dependent member however this increase remains weaker than smoking status and age-related factors and BMI-related risk assessments.

These research findings deliver valuable knowledge to three stakeholder groups including individuals who purchase insurance, insurance provider executives and legislative representatives. Insurance firms can enhance price transparency and equality through the application of these obtained data elements to their pricing models. These study results enable policy decision-makers to maintain healthcare premiums while facilitating lower-cost medical care availability. You can access this information for cost savings to understand which behaviors and traits modify your insurance policy rates.

The research demonstrates how decision-making based on data stands vital to the healthcare insurance sector. Decisions made by the healthcare insurance industry require data-based assessments according to this research. Statistical analysis and empirical data strengthen the ongoing efforts to achieve affordable predictable health insurance.

## **Introduction**

### **Background**

Healthcare insurance acts as an essential safety mechanism to assist people and their families with medical payment expenses (van den Broek-Altenburg & Atherly, 2019). Health insurance costs in multiple countries repeatedly change because of demographical features together with lifestyle elements and geographic locations. Seemingly many insurance policyholders struggle to understand why their insurance payments differ significantly from those of others. Openness is lacking within the healthcare insurance sector which generates issues about financial accessibility alongside concerns for affordability and fairness (Singh et al., 2023).

The complexity of reasons determining insurance rates remains extensive but health services cost increases have turned insurance into a fundamental necessity. Understanding which elements determine insurance premiums becomes possible through comprehension of age, gender, BMI, smoking status, dependent count and geographic location impacts on rates. The research analyzes authentic data to understand effects which influence insurance rates.

### **Problem Statement**

This study focuses on resolving the primary problem related to healthcare insurance cost determinants which currently remain unclear. People in general remain unaware about how their characteristics and behaviors specifically influence premium costs even though insurance companies base their pricing on risk models. The standard policy information systems do not explain how body mass index relates to premium rates. The general public remains unaware of the reason why smokers must pay higher insurance costs.

Premiums in health insurance show variations based on where people reside in the country. Some areas experience higher healthcare costs because their residents have different levels of medical care access together with specific provider billing strategies in addition to payment rates for regional treatments. The proposed research aims to investigate the varying insurance rates so it can contribute empirical findings about health insurance pricing.

### **Research Objectives**

This study aims to address these issues by concentrating on the following goals:

1. The research focuses on achieving the subsequent objectives to resolve these problems.

- 2.The analysis will assess how age combined with gender and dependent count influences the insurance price.
- 3.The analysis examines how both smoking status together with body mass index determine the price of insurance coverage.
- 4.This study investigates reasons behind varying insurance premium costs that occur across geographical areas.
- 5.This study provides data to help create healthcare policies with fair distribution as well as optimize insurance pricing methods.

### **Significance of the Study**

Multiple groups derive crucial value from the research findings:

1. The collection of health-related data by insurance companies permits them to establish pricing systems that align premiums with genuine medical risks at approachable rates.
2. Policymakers along with regulatory agencies can implement healthcare insurance policy modifications through the results to manage premiums and eliminate unfair populational discrimination.
3. Insights into which factors determine insurance rates enable people and policyholders to create better decisions about their financial planning and healthcare needs.

### **Scope of the study**

The insurance data collection for this research includes seven features which comprise age, gender and BMI statistics along with smoking habits and area information as well as dependents and insurance fee metrics. The data analysis combines statistical methods with exploratory data analysis to establish the factors influencing insurance rates. This research focuses solely on analyzing variables from the provided dataset because outside elements affect insurance costs although they are not part of this study.

### **Conclusion**

This study aims to decrease the understanding difference between insurance pricing methods and public comprehension by revealing essential variables that shape health insurance premiums. Statistical analysis alongside data analytics develops a fairer healthcare insurance market which becomes more transparent.

## LITERATURE REVIEW

The chosen dataset "Healthcare Insurance", from Kaggle, has been a base for many research jobs within the healthcare domain. Introduced here are five of the most important studies along with their results and why and how this project will provide value addition:

1. **Health Insurance Fraud Detection Using Data Mining:** This study deals with the ongoing issue of fraud in healthcare insurance. The authors elaborated on the difficulties met in the identification of new and sophisticated ways of committing fraud, thus highlighting the necessity of using effective data analytic and advanced techniques to identify any fraudulent activities within the health insurance claims.
2. **Healthcare Cost Pattern and Prediction Studies:** A Data analysis approach in this research, healthcare costs patterns are analyzed by means of personal datasets. Data analytics is used, leading to some findings and healthcare expenditures being predicted in the future, to shed more light on cost drivers in the health system.
3. **Comparative Analysis to Predict Medical Health Insurance Cost Using Machine Learning Algorithms:** This analysis compares Lasso regression, Ridge regression, KNN, and XGBoost to predict medical health insurance costs. The study proved more predictive accuracy using XGBoost in comparison with others, thus establishing its high credibility for medical cost predictions (Currie et al., 2019).
4. **Analysing Health Insurance Customer Dataset to Determine Cross-Selling Opportunities Using Machine Learning Algorithms:** In this research, machine learning algorithms were used in determining possible cross-selling opportunities in health insurance based on customer datasets. With regard to customer demographics and behavior mining, the analysis aims to assist insurance companies in planning for targeted marketing campaigns and personalized offers to turn goals into revenue.
5. **Health Insurance Data Analysis and Visualization:** Provides a statistical analysis and visualization of health insurance charges with R. The study explores several major factors- age, BMI, smoking status, and region-by employing regression models and visualizations, providing insight to further enhance affordability, accessibility, and equity in healthcare coverage.
6. **Fraud Detection in Healthcare Insurance Claims Using Machine Learning:** This work outlines the use of supervised machine learning algorithms like random forests and logistic



regression for the detection of fraudulent claims in health insurance. The study outlined the working of these models in forecasting fraud in voluminous data to lessen the financial impact of fraudulent activities(Alhassan, Adetiba, & Ojo, 2024).

7. Health Insurance Premium Prediction With the Application of Machine Learning: This study focuses on using learning algorithms based on regression to predict the premiums for health insurance. The study would analyze the individual characteristics to estimate the costs accurately and thus help insurers in pricing and financial planning (Currie et al., 2019).
8. Using Interpretable Machine Learning Methods: An Application to Health Insurance Fraud Detection: This paper looks into the application of interpretable machine learning methods in the problem of health insurance fraud detection. The study aims at balancing the classes of fraudulent and nonfraudulent cases and offers some solutions meant to bolster reliability and transparency of the models.
9. Fraudulent Health Insurance Claims Detection Using Machine Learning: This research studies the application of some artificial intelligence models like multi-layer perceptron neural networks in conceiving ways to detect fraudulent health insurance claims. A monthly detection rate of 75 frauds is reported, which further highlights the power of neural networks in highlighting complex fraud behavior.
10. Machine Learning for the Explainable Prediction of Medical Insurance Costs: Ensemble machine learning models, such as Extreme Gradient Boosting and Random Forest, are integrated into this study to predict the costs of medical insurance. XAI methods, including Shapley Additive Explanations (SHAP) and Individualized Conditional Expectation (ICE) plots), enable transparency by clarifying the important causes of insurance premium specification. Disallow grammatical and word errors.

### **Differentiation of Contribution:**

This project may uniquely illuminate the integration of audit detection, cost prediction, and cross-selling opportunities, which up to now have been rather unrehearsed with machine learning and de Christensen.

Possible avenues for novel contributions include:

Integrated Treatment: Integrate fraud detection, cost prediction, and customer segmentation into a single model for a broader view on healthcare insurance.

Upgrade in Real Time: Enable real-time capability to process data and detect fraud.

## Data

The "Healthcare Insurance" dataset at Kaggle contains important data points to understand the different elements which affect healthcare insurance rates(Gibin, n.d.). This database consists of multiple features which include demographic information on age together with gender attributes as well as BMI measurements and family child counts alongside tobacco usage records and geographical records. The insurance charge constitutes the target variable in the dataset because it shows the payment amount for medical coverage. The dataset provides an excellent opportunity to analyze insurance price generation because it contains comprehensive variables about individual demographics along with lifestyle and health-related data.

Among all the dataset features age stands out since older people face higher insurance premiums because they have a higher susceptibility to healthcare complications. Premiums rise for individuals who smoke because the health risks from smoking increase their rate of insurance coverage. The insurance premium depends on BMI which stands for Body Mass Index as this health measure proves significant when insurers set rates. Insurance customers with elevated Body Mass Index commonly face higher premiums since their increased health risks increase the chance of developing diseases like heart disease and diabetes. The insurance premium increases when policyholders have multiple children because they need to include their children under the coverage. The premiums of insurance plans often depend on the insured individual's sex although insurers and geographic locations enforce different levels of influence.

Several data science techniques enable analysis of the dataset beginning with Exploratory Data Analysis (EDA) as the base approach. EDA provides the opportunity to examine variable relationships and visualize the distribution patterns through scatter plots and other graphical models such as histograms and box plots for insurance charge data against features. A scatter plot presents evidence for positive relationship between age and insurance cost where insurance premiums grow in line with subject age rise. Bar charts would effectively show the relationship between region-based average premiums because location factors significantly impact health insurance pricing.

The existing dataset matches perfectly with regression analysis models like linear regression for making insurance charge estimates based on available features. Historical data allows the model to understand how independent variables relating to age, sex, BMI, smoking status, region and number of children impact the dependent variable which is insurance charge. The trained models enable predictions of upcoming medical premiums for fresh patient groups from their specific features.

The dataset serves as an exceptional learning tool for people who want to build their data analysis and machine learning competencies. In educational settings the dataset serves as material to teach students how to process real datasets and clean them while learning different machine learning methods. Data analysis for this dataset becomes more effective because of Python tools particularly Pandas and Matplotlib and Seaborn.

The "Healthcare Insurance" dataset hosted by Kaggle represents an extensive resource about healthcare premium determinants for all those studying this topic. First-time students and expert data scientists will find plenty of educational prospects within this dataset to develop their data analysis abilities together with machine learning and predictive modeling competencies. The dataset is available for direct work on Kaggle through this link.

### Source of data

- Source: **Kaggle Main Website:** <https://www.kaggle.com>

Below is a detailed summary of the Healthcare Insurance dataset, including its rows, columns, features, and target variable:

- **No of Rows:** 1338
- **No of Columns:** 7

Column Name	Data Type	Description	Role
<b>age</b>	Integer	Age of the policy holder	Feature
<b>sex</b>	Categorical (String)	Gender of the individual (male/female)	Feature
<b>Bmi</b>	Float	Body Mass Index, a measure of body fat based on height and weight	Feature
<b>children</b>	Integer	Number of children covered by the insurance policy	Feature
<b>smoker</b>	Categorical (String)	Indicates whether the individual is a smoker (yes/no)	Feature
<b>region</b>	Categorical (String)	Geographic area of residence (e.g., northeast, southeast, etc.)	Feature
<b>charges</b>	Float	The medical insurance cost billed to the individual (in dollars)	Target

The dataset serves many predictive modeling purposes mainly through regression analysis that uses various features to forecast insurance premium (charges). Researchers analyze the insurance costs by studying how age and BMI alongside smoking status behaviors interact with each other. The dataset possesses numerical along with categorical variables that makes it appropriate for multiple preprocessing techniques including variable normalization and categorization processes.

Three different visual methods exist for data representation purposes.

Analyzing continuous variables with the target variable (charges) works best through the use of scatter plots.

The feature distributions become transparent through the use of histograms when analyzing BMI or age patterns.

The analysis utilizes bar charts to display mean charges based on sex and region classification variables.

Correlation Matrices: These provide a visual representation of the strength and direction of relationships between different variables.

An extensive database has been obtained from the Kaggle platform through Willian Oliveira Gibin who made it available at <https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance>. The expansive information set functions as both a valuable educational tool for data cleaning practices as well as exploratory data analysis and visualization techniques and simultaneously serves as a strong basis for building healthcare economics prediction models.

## **Methods**

We use Exploratory Data Analysis (EDA) together with statistical tests as well as machine learning techniques to study the dataset for the derivation of significant insights. This project seeks to discover major insurance charge influence variables before creating predictive models to sort charges into Low, Medium and High sections. The implementation methods for this project consist are:

### **Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) refers to a process or methodology of summarizing the main characteristics of a dataset with visual methods. EDA is a must first step of data analysis which helps the analyst to understand the data structure, detect patterns, detect outliers and test hypothesis before applying more formal statistical techniques. The following techniques were used in this project:

- **Data loading and Initial Exploration:** We loaded the dataset into R using the `read.csv()` function. Functions like `str()`, `dim()`, `head()`, and `tail()` were used to do basic exploration of the dataset to understand what the structure and content of the dataset is. To run a detailed summary of the dataset using the `skim()` function in the `skimr` package, this included missing values, data types, and descriptive statistics.

```

{r}
my_data <- read.csv("insurance.csv")
my_data
str(my_data)
dim(my_data)
head(my_data)
tail(my_data)

```

```

'data.frame': 1338 obs. of 9 variables:
 $ age      : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex      : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi      : num  27.9 33.8 33 22.7 28.9 ...
 $ children : int   0 1 3 0 0 0 1 3 2 0 ...
 $ smoker   : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region   : Factor w/ 4 levels "northeast","northwest",...: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges  : num  16885 1726 4449 21984 3867 ...
 $ bmi_category : Factor w/ 4 levels "Normal","Obese",...: 3 2 2 1 3 3 2 3 3 3 ...
 $ charge_category: Factor w/ 3 levels "High","Low","Medium": 3 2 2 3 2 2 2 2 2 3 ...

```

Fig 1: Data Exploration

- Data Visualization: Here we use several visualizations of the distribution of variables were made and relationship between them:
  - Use bar plots for visualize distribution of such categorical variables (sex, smoker, region etc.)
  - Use histogram for analysis of numerical variable such as age, bmi, children and target variable charges to see if its skewed and what is the range.
  - Use boxplots to detect and analyze outliers in our dataset (like bmi, charges etc.)
  - Using correlation heatmaps to find relation between numerical variables (age, bmi, children, charges).
- Feature Engineering: In order to simplify the analysis and increase the model performance, we created new variables in our project. We defined BMI categories such that it falls under 4 groups (Underweight, Normal, Overweight, Obese). To simplify the prediction task, Charge categories variables were also invented as the three group variables (Low, Medium, High).

## Statistical Data Analysis Methods

The statistical data analysis methods are developed to summarize, interpret and draw sensible conclusion from the data. In this study, the following statistical techniques are used:

- Descriptive Statistics: This technique is used to summarize the numerical and categorical data. Mean, median, standard deviation and mode are computed as numerical variables. Frequency counts are provided for categorical variables.
- Statistical testing of hypotheses: The second step of finding the significant relationships between variables and charges is statistical testing of hypotheses. They are:

- T-tests: Used to compare mean charges between two groups (smokers vs. non-smokers, males vs. females).
- ANOVA: Here in this project we use anova to test the impact of categorical variables i.e, BMI categories and region on charges.
- Correlation Analysis: In this, we generated a correlation heatmap to get relationship between numerical variables (age, bmi, children, charges).

## Data Cleaning and Preprocessing

Several steps were taken in the data cleaning and preprocessing phase so that the dataset was ready for analysis. For missing values, imputation or removal was done to have no gaps in the data. Id columns were dropped with main features to make the dataset simpler such that our focus is on the relevant features. To prevent key variables having outliers (e.g. BMI, charges) skewing the analysis, they were identified and the potential outlier cases were dealt with to achieve a cleaner dataset with a lower noise level that facilitate further exploration.

## Statistical and Machine Learning Methods

Using statistical and machine learning methods, we will correctly analyse and determine the interrelation between dependent variable and other independent variables. We will also use these methods with application, validation and a better understanding of how accurate a prediction can be made on our insurance dataset. The model used are:

- a. **Linear Regression:** The linear regression is a statistical method to model the relationship between dependent variable and one or more independent variables. The goal is to fit a linear equation to observed data in order to predict the dependent variable as a linear function of the independent variables. In our case, we had a linear regression model predicting insurance charges based on the factors like age, BMI, smoking status, number of children, and region. The most important variables concerning charges were smoking status and age, which explain the most importance in setting insurance costs.
- b. **Decision Tree:** It is supervised machine learning model which splits the data in subsets of input features to form a tree like structure. A decision is represented by each node and an outcome by each leaf. Decision trees are easy to interpret, but easy fit. In our

case, the model was good in predicting Low and High but not Medium charges possibly because of data imbalance.

- c. **Random Forest:** A random forest is a technique where we combine many decision trees. The output produced by random forest is the output which is selected by most of its trees for classification. Logistic regression is a slightly less complex model than this one to implement and interpret than decision tree. This method gives better model robustness, higher accuracy and prevents overfitting problems.

## Information on Data Analysis Process

The data analysis process followed these steps:

- a. **Data Collection:** The dataset was loaded and inspected.
- b. **Exploratory Data Analysis (EDA):** The dataset was used to create visualizations and statistical summaries to better understand it.
- c. **Data Cleaning and Preprocessing:** Imputation and removal were employed to handle missing values and unknown values.
- d. **Feature Engineering:** Features not needed were removed and new features were created to further enhance model performance.
- e. **Model Development:** Finally, all the models had been trained and evaluated with Linear Regression, Decision Tree, and Random Forest model.
- f. **Model Evaluation:** Accuracy, precision, recall, F1-score and other metrics were used to evaluate the models.

## Relationship Between Methods and Research Objectives

The research objectives are satisfied by the methods used in this project is:

- EDA and Correlation Analysis to identify the key factors that influence the insurance charges and how age, bmi, and smoker status affects the charges.
- Statistical and Machine Learning Models: Understand how well charge categories can be predicted to set more appropriate and accurate premiums for insurers along with risk management.

## Results and Discussion

The analysis of the insurance.csv dataset, containing 1,338 records of individuals' demographic and lifestyle factors and corresponding medical charges is presented in this section. The analysis is broken into three parts they are:

- A. Exploratory Data Analysis
- B. Statical Analysis
- C. Predictive Modeling

### Exploratory Data Analysis

The first step in understanding the dataset is performing Exploratory Data Analysis (EDA). Summarizing the main characteristics of the data, with sometimes using visual method. In this section, there are explored distribution of categorical, and numerical variables, and the relationship between them (Lin & Chen, 2024).

Here are the data distribution we can see them by plotting bar chart for categorical variable and histogram for numeric variables.

#### Categorical Variables

Distribution of Gender

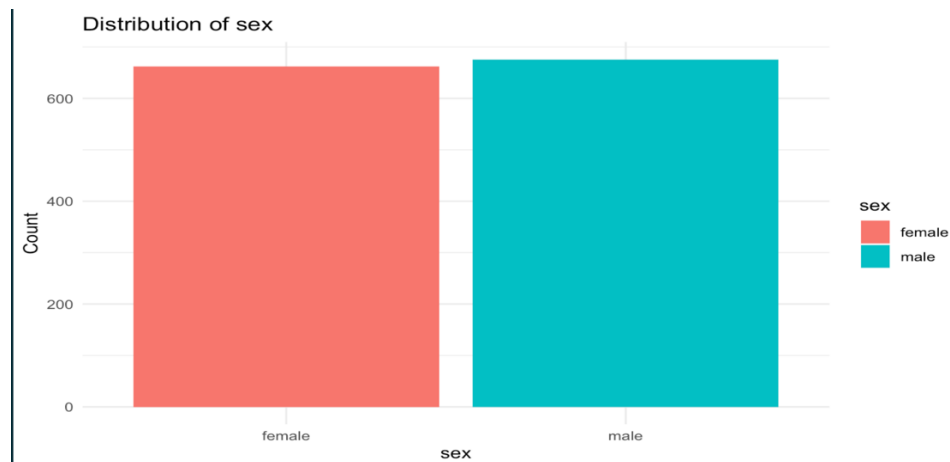


Fig 2: Distribution of Gender

Insight: The bar chart shows the count of the males and females within a dataset. The chart reveals that the number of males and females is almost close and there is a very small difference on the number of males over females, which seems to be a little more. This may



indicate that the dataset treats both genders equally, and so does not have any gender bias in analysis.

### Distribution of Smoker

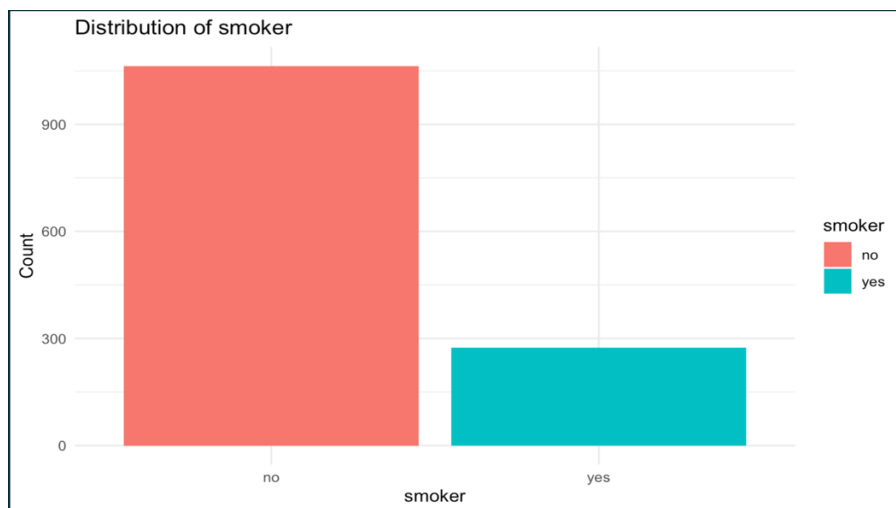


Fig 3: Distribution of Smoker variable

Insight: In the bar chart, the distribution of smokers and non-smokers is shown in the dataset. The clearly shown visualization shows that most of the people are non-smokers, with much more people in comparison to the smokers.

### Distribution of Region

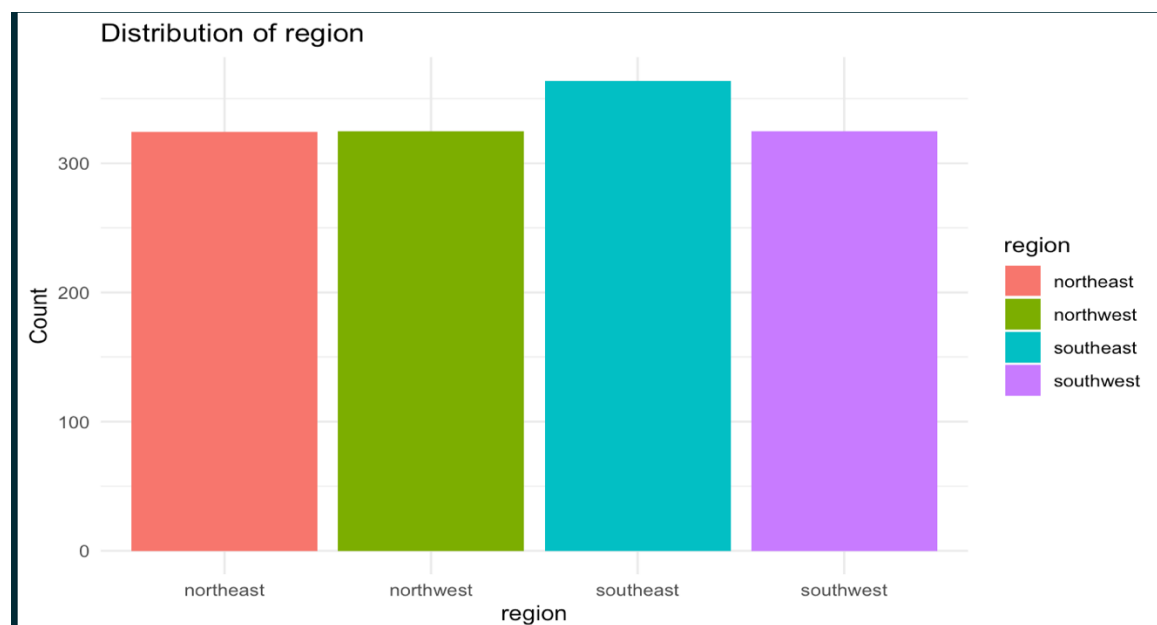


Fig 4: Distribution of Region variable

Insight: A bar chart is shown depicting the distribution of individuals in four regions, namely northeast, northwest, southeast, and southwest. The counts of each region are not far from balance, most of the southeast with the highest count and the other three regions with similar counts. A slight variation in the southeast may indicate there is more population or more data from that area.

## Numerical Variables and Insights

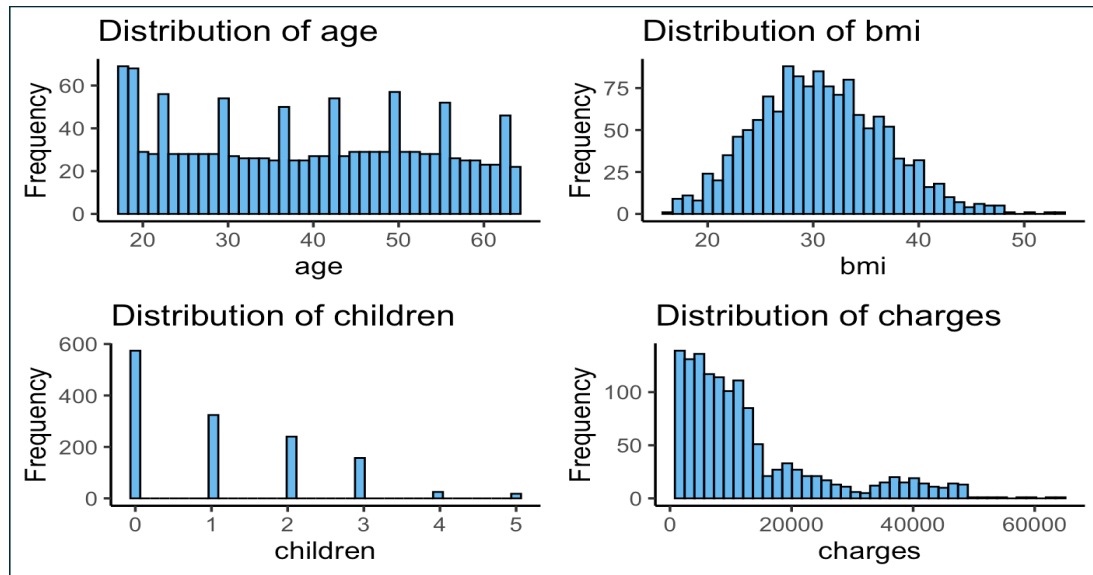
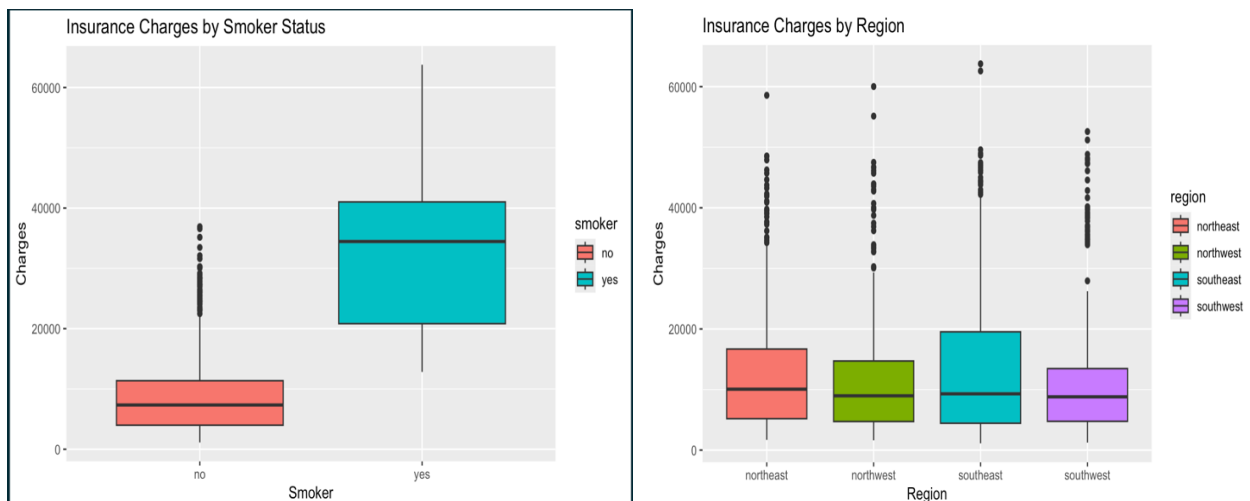


Fig 5: Distributions of Numerical variables

- **Distribution of Age**
  - It seems that age is distributed uniformly, so there roughly equal distribution of age (from 18 to 64 years) in the data.
  - This uniformity is useful for analyzing how medical charges change with age since medical charges are not biased towards any one age group.
  - In addition, the uniform distribution also indicates that the dataset is suitable for age based segmentation and analysis.
- **Distribution of BMI**
  - It appears that the distribution of BMI is roughly normal, with most people having a BMI between 25 and 35.

- It is possible that the moderate positive correlation between BMI and medical charges is because the majority of individuals are either overweight or obese.
- BMI distribution is normal like and hence it is a reliable variable for analysis as it is not strongly skewed.
- Distribution of Children
  - Children are distributed right skewed, most individuals have 0–2 children. Only a very few have more than 3 children.
- Distribution of Charges
  - The distribution of charges is extremely right skewed, with very few people incurring very high medical costs.
  - Most people (most 95%) have relatively low medical cost (less than \$10,000), while a small proportion (top 5%) has responsibility for a large share of total medical cost.

### Boxplot of charges by Smoker, Region, BMI category, Gender



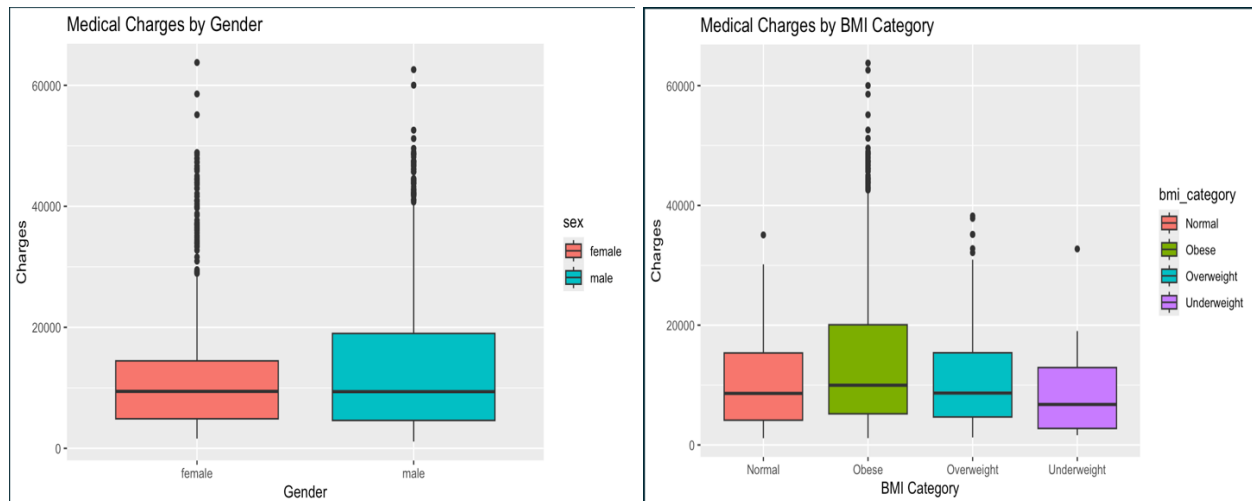


Fig 6: Boxplots of different variables

### Insurance Charges by Smoker Status

- Medical charges for smokers are significantly higher than those for non smokers.
- This shows that the charges of smokers are more variable, so the interquartile range (IQR) is much wider for smokers.
- Some smokers have extremely high medical costs, but there are many outliers in the smoker group they might represent real world data.

### Insurance Charges by Region

- The median charges are relatively similar across all regions.
- The southeast region has slightly higher median charges compared to the other regions.

### Medical Charges by Gender

- The median charges for males and females are similar.
- The IQRs are similar which means the variability in charges is similar among the genders.
- There are some outliers in both groups, but they are not extreme.

### Medical Charges by BMI Category

- Obese individuals have the highest median charges.
- The IQR for obese individuals is wider, indicating greater variability in charges.

In addition, medical charges are significantly higher for smokers, and the southeast's costs are marginally higher. Higher expenses do not depend on gender but rather on the level of BMI, with the most expensive group being obese persons.

### Scatter plot of BMI vs. Charges by Smoker Status

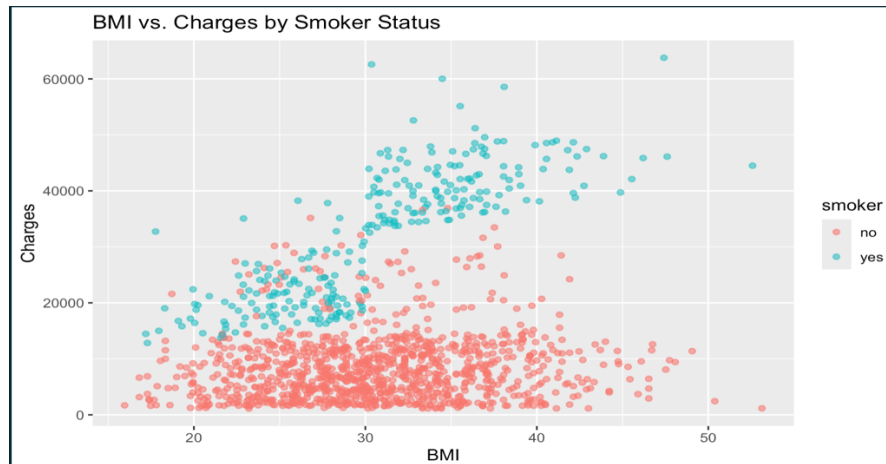


Fig 7: Scatter plot of BMI vs. Charges by Smoker Status

#### Insight:

- Medical charges are significantly higher for smokers regardless of their level of BMI than for non smokers.
- Smoking is a major driver of healthcare costs, and even at lower BMI levels, smokers pay much higher charges.
- Charges go up with higher BMI, particularly in obese people, and non smokers tend to have lower charges.

## Correlation heatmap

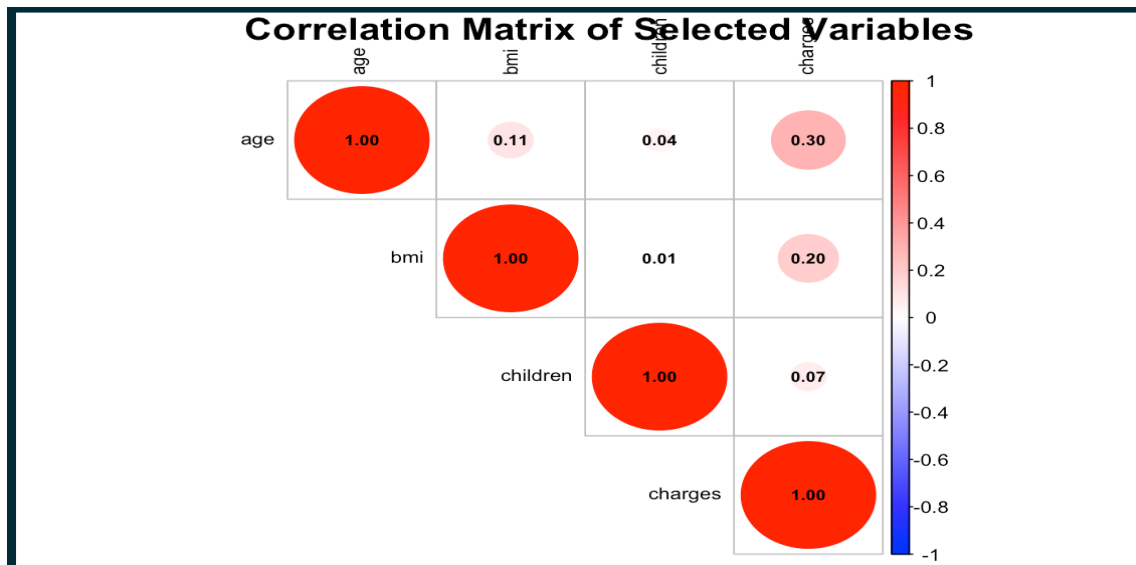


Fig 8: Correlation heatmap

### Insight:

- Age and Charges: Medical charges are higher for older people for reasons of age-related health issues.
- BMI and Charges: Obesity can contribute to higher medical charges, perhaps due to disease such as diabetes and heart disease, which are linked to higher BMI.
- Children and Charges: Medicare charges have little or no relationship with the number of children.

## Statistical Analysis

### T-tests

- T-test for charges by smoker: The idea behind the hypothesis test is to check if there exists a significant difference (Damiati, 2020) in average medical charges between smokers and non smokers. The test statistic ( $t = -32.752$ ), degrees of freedom ( $df = 311.85$ ) indicate a strong deviation from the null hypothesis. This extremely small p-value ( $< 2.2e-16$ ) is strong evidence that smokers and non smokers have different mean charges. The 95% confidence interval ( $-25,034.71$  to  $-22,197.21$ ) supports this difference.

```

Welch Two Sample t-test

data:  charges by smoker
t = -32.752, df = 311.85, p-value < 2.2e-16
alternative hypothesis: true difference in means between
95 percent confidence interval:
 -25034.71 -22197.21
sample estimates:
mean in group no mean in group yes
      8434.268      32050.232

```

Fig 9: T-test for charges by smoker

Mean medical charges for non-smokers are 8,434.27 and for smokers mean is 32,050.23. This implies that smoking results in higher health care costs as smoking related illness. Results show a large difference between charges, emphasizing financial cost of smoking amongst individuals and systems of healthcare.

- T-test for charges by sex: In hypothesis test, we investigate whether there is a significant difference in mean medical charges of females and males. The result gives a test statistic ( $t = -2.1009$ ) and degrees of freedom ( $df = 1313.4$ ) that indicate a small deviation from the null hypothesis. Since p value (0.03584) is less than 0.05, we have sufficient evidence to reject the null hypothesis. The 95% confidence interval (-2682.49 to -91.86) supports the presence of a difference mean charges.

```

Welch Two Sample t-test

data:  charges by sex
t = -2.1009, df = 1313.4, p-value = 0.03584
alternative hypothesis: true difference in means between
95 percent confidence interval:
 -2682.48932 -91.85535
sample estimates:
mean in group female mean in group male
      12569.58      13956.75

```

Fig 10: T-test for charges by sex

The sample estimates of mean charge are 12,569.58 for females and 13,956.75 for males. This implies that males in general spend more on medical expenses than females. Although the difference is statistically significant, it is small in comparison to other factors that might affect healthcare costs.

## ANOVA

- BMI Categories: Results of the ANOVA test indicate that there are significant differences in medical charges based on different BMI categories(Damiati, 2020). At the extremely small p-value ( $6.66 \times 10^{-12}$ ) we have strong evidence against the null hypothesis that BMI has not a large impact on charges.

```

{r}
## ANOVA: Do BMI Categories Affect Medical Charges?
anova_test <- aov(charges ~ bmi_category, data = my_data)
summary(anova_test)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bmi_category	3	7.925e+09	2.642e+09	18.73	6.66e-12 ***
Residuals	1334	1.881e+11	1.410e+08		

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Fig 11: Check BMI significant using ANOVA

This conclusion is reinforced by the F value of 18.73 that shows more variation between BMI categories than within them. This confirms that medical expenses are determined by BMI.

- Region: The test results for ANOVA showed that medical charges vary significantly between regions. Since the p value (0.0309) is less than 0.05, there is enough evidence to reject the null hypothesis that region has no effect on medical costs.



```

{r}
aov_result <- aov(charges ~ region, data = my_data)
summary(aov_result)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	3	1.301e+09	433586560	2.97	0.0309 *
Residuals	1334	1.948e+11	146007093		

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Fig 12: Check Region significant using ANOVA

This finding is supported by the F value of 2.97 which indicates that there are notable differences in charges between regions over variances in within region charges. This shows that location is in play when it comes to paying for medical expenses.

### Linear Regression:

In the case of linear regression, the medical charges are predicted from age, BMI, smoking status, number of children, and region. The model is highly significant with a p.value of less than 2.2e-16 and the R.squared value of 0.7509 explains 75 percent of the variation in charges by predictors. Age, BMI, and smoking status have very strong impacts, but smoking status is the largest. Regions do have some importance, but not as much as their effects are. Overall, the model describes well the relationship between these variables and medical charges.

```

{r}
# Linear Regression
lm_model <- lm(charges ~ age + bmi + smoker + children + region, data = my_data)
summary(lm_model)

```

```

Call:
lm(formula = charges ~ age + bmi + smoker + children + region,
    data = my_data)

Residuals:
    Min       1Q   Median       3Q      Max
-11367.2 -2835.4  -979.7   1361.9 29935.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11990.27    978.76  -12.250 < 2e-16 ***
age           256.97     11.89   21.610 < 2e-16 ***
bmi           338.66     28.56   11.858 < 2e-16 ***
smokeryes    23836.30    411.86   57.875 < 2e-16 ***
children      474.57    137.74    3.445 0.000588 ***
regionnorthwest -352.18    476.12   -0.740 0.459618
regionsoutheast -1034.36    478.54   -2.162 0.030834 *
regionsouthwest -959.37    477.78   -2.008 0.044846 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6060 on 1330 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7496
F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16

```

Fig 13: Linear Regression

## Predictive Modeling

### Data Splitting

The given dataset is split into training and testing sets to test the performance on unseen data, 80% of the training set (1,070 records) and 20% of the testing set (268 records) are used for the below steps. To ensure the split was reproducible, a random seed (123) was applied.

```
```{r}
# Split data into training and testing 80% train and 20% test
set.seed(123)
trainIndex <- createDataPartition(my_data$charge_category, p = 0.8, list = FALSE)
trainData <- my_data[trainIndex, ]
testData <- my_data[-trainIndex, ]
```
```

Fig 14: Data Splitting

- **Decision Tree Model:** Here how we implement the decision tree in our dataset.

```
```{r}
# Decision Tree Classification Model
dt_model <- rpart(charge_category ~ age + bmi + children + smoker + region, data = trainData, method = "class")
pred_dt <- predict(dt_model, testData, type = "class")
```
```

Fig 15: Decision Tree Model

```
```{r}
# Confusion Matrix and Metrics for Decision Tree
cm_dt <- confusionMatrix(pred_dt, testData$charge_category)
print(cm_dt$table)
```
```

|            | Reference |     |        |
|------------|-----------|-----|--------|
| Prediction | High      | Low | Medium |
| High       | 30        | 0   | 0      |
| Low        | 1         | 136 | 12     |
| Medium     | 1         | 6   | 80     |

Fig 16: Confusion Matrix for Decision Tree

- **Random Forest Model:** Here how we implement the random forest in our dataset.

```
```{r}
# Random Forest Classification Model
rf_model <- randomForest(charge_category ~ age + bmi + children + smoker + region, data = trainData, ntree=100)
pred_rf <- predict(rf_model, testData)
```
```

Fig 17: Random Forest Model

```
```{r}
# Confusion Matrix and Metrics for Random Forest
cm_rf <- confusionMatrix(pred_rf, testData$charge_category)
print(cm_rf$table)
```
```

|            | Reference |     |        |
|------------|-----------|-----|--------|
| Prediction | High      | Low | Medium |
| High       | 30        | 0   | 0      |
| Low        | 1         | 138 | 13     |
| Medium     | 1         | 4   | 79     |

Fig 18: Confusion Matrix for Random Forest

To validate our model we use confusion matrix, F1-score, precision, accuracy, recall. Here they are:

### Confusion matrix

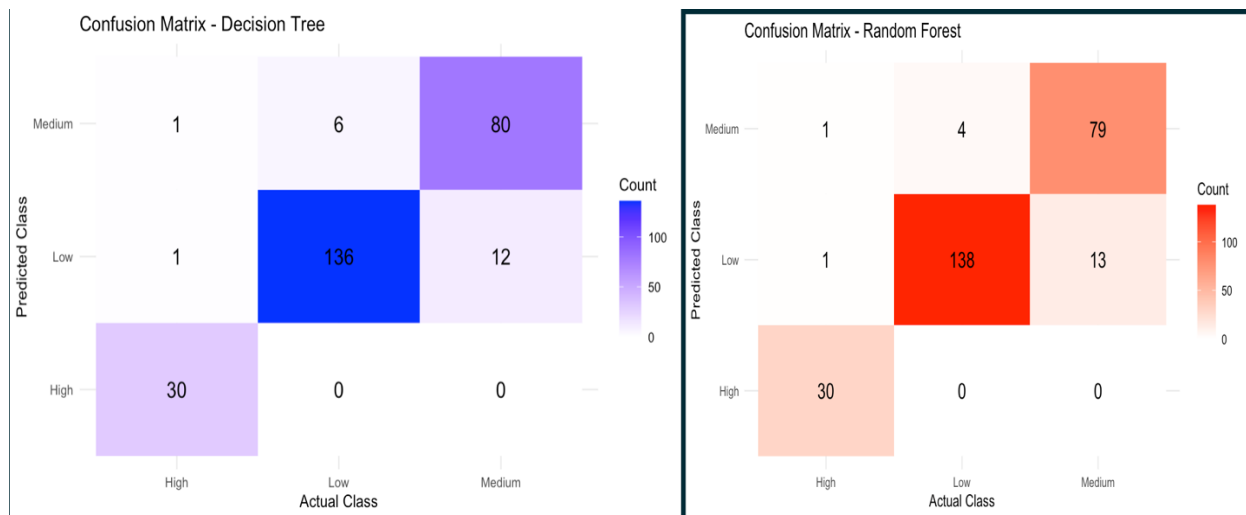


Fig 19: Visualization of Confusion Matrices

The confusion matrix obtained from the Random Forest model indicates how accurately it predicted the classes (Low, Medium, High). This correctly classified 138 "High" instances and 79 "Low" instances and misclassified 13 "Low" instances to "Medium." The model predicts 'High' and 'Low' classes very well, but a little less well for the class of 'Low', with respect to 'Medium'.

The same result was produced by the Decision Tree model, as well as with slightly lower accuracy. It successfully classified 136 "High" and 80 "Low" examples correctly, but misclassified 12 "Low" examples as "Medium" and 50 "Medium" examples as "Low." Both models give reasonable predictions for the "High" charges and fail more in predicting "Low" and "Medium" charges which may suggest that tuning or feature engineering can potentially help.

### **F1-score, precision, recall**

Here we show the metrics for each class. The Decision Tree and Random Forest models had very good accuracy on the 'High' class, with perfect precision and high recall. The Decision Tree performs a little better than Random Forest for the "Low" class, and has higher recall and a better F1 score. Using the Random Forest in the "Medium" class, precision and F1 score also get slightly improved over the Decision Tree. Overall, the Random Forest model performs better in the recall for the "Low" class and precision for the "Medium" class, indicating a better classification to perform for this classification task.

| Decision Tree Metrics:      | Random Forest Metrics:      |
|-----------------------------|-----------------------------|
| Metrics for 'High' class:   | Metrics for 'High' class:   |
| Precision: 1                | Precision: 1                |
| Recall: 0.9375              | Recall: 0.9375              |
| F1 Score: 0.9677419         | F1 Score: 0.9677419         |
| Metrics for 'Low' class:    | Metrics for 'Low' class:    |
| Precision: 0.9127517        | Precision: 0.9078947        |
| Recall: 0.9577465           | Recall: 0.971831            |
| F1 Score: 0.9347079         | F1 Score: 0.9387755         |
| Metrics for 'Medium' class: | Metrics for 'Medium' class: |
| Precision: 0.9195402        | Precision: 0.9404762        |
| Recall: 0.8695652           | Recall: 0.8586957           |
| F1 Score: 0.8938547         | F1 Score: 0.8977273         |

Fig 20: F1-score, Recall, Precision of both models

## Accuracy

|   |   |
|---|---|
| <pre> {r} #accuracy for decision tree accuracy_dt &lt;- cm_dt\$overall["Accuracy"] cat("Accuracy:", accuracy_dt, "\n") </pre> | <pre> {r} #accuracy for random forest accuracy_rf &lt;- cm_rf\$overall["Accuracy"] cat("Accuracy:", accuracy_rf, "\n") </pre> |
| Accuracy: 0.924812  | Accuracy: 0.9285714   |

Fig 21: Accuracy of both models

Here the accuracy for both model is almost same which is quit good i.e. 92%.

## Critical Analysis

The study provides useful information for insurers with respect to what factors impact medical charges (e.g. smoking, BMI). The two strategies that may lower costs include targeting smoking

cessation programs as well as promoting healthier BMI levels. Moreover, the conjoining of statistical testing with machine learning models increases the trustworthiness of the analysis.

Although, the study is limited by a small dataset, and therefore the generalizability of the results may be questionable. In linear regression, there is an assumption of a linear relationship which might miss out complex interaction in the data. However, these limitations have important implications for the theory of the pricing model of insurers as well as for policyholders to make informed decisions when making health decisions to reduce medical expenses.

## **Related Questions**

1. What is influence of smoking status on medical insurance charges?  
⇒ Increased risk of serious health conditions such as lung cancer, heart disease and respiratory illnesses leads to higher medical charges among smokers compared to non-smokers who are not smokers. That difference is statistically significant ( $p < 0.05$ ), as confirmed by a t-test, and smoking is a major factor affecting medical expenses.
2. Is there a relationship between age and medical insurance charges?  
⇒ Yes, medical charges have a positive correlation ( $r = 0.30$ ) with age and more costs with age, as the age is related to health problems. The found correlation is statistically significant ( $p < 0.05$ ).
3. What is the influence of BMI on medical insurance charges?  
⇒ BMI is moderately ( $r = 0.20$ ) positively correlated with medical charges, with the more obese (overweight or obese) people having higher costs due to obesity related conditions, such as heart disease, diabetes and joint problems. The differences in medical charges among BMI categories are statistically significant ( $p < 0.05$ ) as confirmed by ANOVA.

## **Limitations and Future Study**

- The accuracy and reliability of information: The condition of the input data has effects on the evaluation correctness unfair outcomes may result error also may result missing data also differences information can be age or other can be error or wrong this outcome might not apply for the vast majority if some groups are missing or error.

- Connection against leakage: The research fails to connection if might include data for instance lifespan and medical costs uncertainties and additional uncertainties may effect association shown as well as produced false incorrect assumptions regarding relevance of particular number.
- Design Overestimation: Overestimation which the prediction functions effectively on information lesson learned on unknown data which is unfamiliar problems which may rise intricate design utilized lacking enough information whether compare freshly created categories of patients this may lead misleading estimations
- The ever-changing circumstance surrounding medicinal treatment: The price of regulations as well as norms regarding or concerning machine are always changing methods laws are developed system that have been built earlier previous information might turn useless because of this the design must update time to time.
- All potential relationship across various factors are not detected by education if patient falls down into a specific population organization such as sex, Marital status or consuming history which has impact on medical expenses.

## Conclusion

In this case study we find the factor that impacting of the health or medical coverage we study about health care information throughout the case study we specifically examined characteristics including Gender, Age, and Body mass index among of youngsters also uses of cigarette usage along with location. Then we also found significant driver of medical health care expenses through put our investigation along with analytical framework showing illustrated or connections among above all of this all of expenses paid by insurance people.

Although overall insightful discoveries a number of drawbacks emerged such as problems within information data accuracy the complexity proving connection possible excessive modeling as well as the constantly changing character of Health care further more integrity the conclusions may effectively by Body Mass Index information reduction complicated relationship

The primary findings and throughout research highlight significance using outstanding knowledge and the requirement to maintain models.



# Appendix

## References

- Alhassan, G. N., Adetiba, E., & Ojo, J. (2024). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*, 15, 100516. <https://doi.org/10.1016/j.mlwa.2023.100516>
- Cevolini, A., & Esposito, E. (2020). From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data & Society*, 7(2). <https://doi.org/10.1177/2053951720939228>
- Currie, G., Hawk, K. E., Rohren, E., Vial, A., & Klein, R. (2019). Machine learning and deep learning in medical imaging: Intelligent imaging. *Journal of Medical Imaging and Radiation Sciences*, 50(4), 477–487. <https://doi.org/10.1016/j.jmir.2019.09.005>
- Damiati, S. A. (2020). Digital pharmaceutical sciences. *AAPS PharmSciTech*, 21(6), 206. <https://doi.org/10.1208/s12249-020-01747-4>
- Gibin, W. O. (n.d.). *Healthcare Insurance Dataset*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance>
- Hanafy, M., & Mahmoud, O. M. A. (2021). Predict health insurance cost by using machine learning and DNN regression models. *International Journal of Innovative Technology and Exploring Engineering*, 10(3), 137–143. <https://doi.org/10.35940/ijitee.C8364.0110321>
- Lin, C. P., & Chen, L. A. (2024). Application of artificial intelligence models in nursing research. *Hu Li Za Zhi*, 71(5), 14–20. [https://doi.org/10.6224/JN.202410\\_71\(5\).03](https://doi.org/10.6224/JN.202410_71(5).03)
- Singh, Y. R., Shah, D. B., Kulkarni, M., Patel, S. R., Maheshwari, D. G., Shah, J. S., & Shah, S. (2023). Current trends in chromatographic prediction using artificial intelligence. *Analytical Methods*, 15(23), 2785–2797. <https://doi.org/10.1039/d3ay00362k>
- ul Hassan, C. A., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A. A., & Sajid Ullah, S. (2021). A computational intelligence approach for predicting medical insurance cost. *Mathematical Problems in Engineering*, 2021, Article ID 1162553. <https://doi.org/10.1155/2021/1162553>
- van den Broek-Altenburg, E. M., & Atherly, A. J. (2019). Using social media to identify consumers' sentiments towards health insurance. *Applied Sciences*, 9(10), 2035. <https://doi.org/10.3390/app9102035>
- Wichmann, J., & Eberl, S. (2022). Deep learning for prediction of population health costs. *BMC Medical Informatics and Decision Making*, 22(1), 32. <https://doi.org/10.1186/s12911-021-01743-z>