“Dissemination Education for Knowledge, Science and Culture”

-Shikhanmaharshi Dr. Bapuji Salunkhe

(स्वायत्त) कोल्हापूर

VIVEKANAND COLLEGE, KOLHAPUR

(Empowered Autonomous)

DEPARTMENT OF STATISTICS

A PROJECT REPORT

On

“STATISTICAL ANALYSIS ON HEALTHCARE INSURANCE”

Submitted by

Ms. Patil Priti Babasaheb

Ms. Ekal Vedika Shivanand

In partial fulfillment for the award of

the degree of

MASTER OF SCIENCE

In

STATISTICS

2023-24

CERTIFICATE

This is to Certify that,

Sr. No.	Name	Roll No.
1	Ms. Patil Priti Babasaheb	1418
2	Ms. Ekal Vedika Shivanand	1406

Have satisfactorily completed the project work on **"Statistical Analysis On Healthcare Insurance"** as a part of practical evaluation for **M.Sc. II,** prescribed by the Department of Statistics, ***Vivekanand College, Kolhapur (Empowered Autonomous)*** in the academic year **2023-24.**

This project has been completed under our guidance and supervision. To the best of our knowledge and belief, the matter presented in this project report is original and has not been submitted elsewhere for any other purpose.

29/05/24
Project Guide
(Ms. Patil R.M.)

29/05/24
Examiner

Head
(Ms. Pawar V. V.)

HEAD
DEPARTMENT OF STATISTICS
VIVEKANAND COLLEGE, KOLHAPUR
(EMPOWERED AUTONOMOUS)

ACKNOWLEDGEMENT

As we approach the culmination of this enormous endeavor, it brings immense satisfication to reflect upon and express gratitude for the collective efforts and support from all those who enthusiastically contributed. Together, we have transformed what once seemed like afar-off aspiration for industrial training into a reality.

We are thankful to Principal Dr. R. R. Kumbhar for his kind cooperation and valuable support. We are also thankful to Mrs.V.C.Shinde, Head ,Department of Statistics. VivekanandCollege, Kolhapur (Empowered Autonomous). We are indeed grateful to OJT Coordinator Prof. R.M.Patil Madam, Assistant Professor, Departrnent of Statistics, Vivekanand College, Kolhapur (EmpoweredAutonomous) for his kind and valuable support.

We are extremely thankful to project guide Prof. R.M.Patil Madam. Assistant Professor, Department of Statistics, Vivekanand Colleges Kolhapur (Empowered Autonomus) for his valuable guidance and mentorship throughout this project work given to us during the study.We are also thankful toall the teaching and non-teaching staff of our department for their direct and indirect support. We also thanks to all the members of Next Level Design Academy. Additionally, we sincerely thanks to our parents, friends and colleagues for their directand indirect contributions.

Yours sincerely,
M.Sc. II
Department of Statistics

INDEX

Sr.No.	Content	Page No.
1	Introduction	5
2	Scope of Study	6
3	Objectives	7
4	Data Collection	8
5	Methodology	9
6	Graphical Representation	10
7	Statistical Analysis	14
8	Conclusion	21
9	Reference	22

INRODUCTION

Health insurance is a critical component of personal financial planning and healthcare access, often influenced by a variety of demographic and lifestyle factors. This project aims to explore and analyze the relationships between different factors such as age, sex, Body Mass Index (BMI), number of children, smoking status, geographic region, and insurance charges. In recent years, healthcare costs have surged, making it increasingly important to understand the factors influencing medical expenses. Health insurance premiums and out-of-pocket costs are critical components of financial planning for individuals and families.

This project seeks to explore the determinants of medical charges using a dataset that includes various attributes such as age, sex, body mass index (BMI), number of children, smoking status, and geographic region. By analyzing these variables, we aim to uncover patterns and relationships that can inform policy makers, insurers, and healthcare providers.

The dataset under consideration comprises detailed information on ten individuals, highlighting a range of profiles from different demographics and regions.

SCOPE OF STUDY

Understanding the factors that drive healthcare costs is essential for developing effective health policies and insurance strategies.

This project will provide insights into how different variables contribute to medical expenses, enabling stakeholders to make informed decisions. Additionally, the findings could help in designing targeted interventions to reduce healthcare costs and improve population health outcomes

OBJECTIVES

- To identify key factors affecting on medical charges.
- To investigate the relationship between BMI and medical charges.
- To examine the impact of smoking on medical expenses.
- To analyze demographic influences (age, sex, and number of children) on charges.
- To assess regional variations in medical charges.

DATA DISCRIPTION

Data Source:

Dataset is from the kaggle site. This dataset contains information on the relationship between personal attributes (age, gender, BMI, family size, smoking habits), geographical factors, and their impact on medical charges. It can be used to study how these features influence insurance costs & develop predictive models for estimating healthcare expenses.

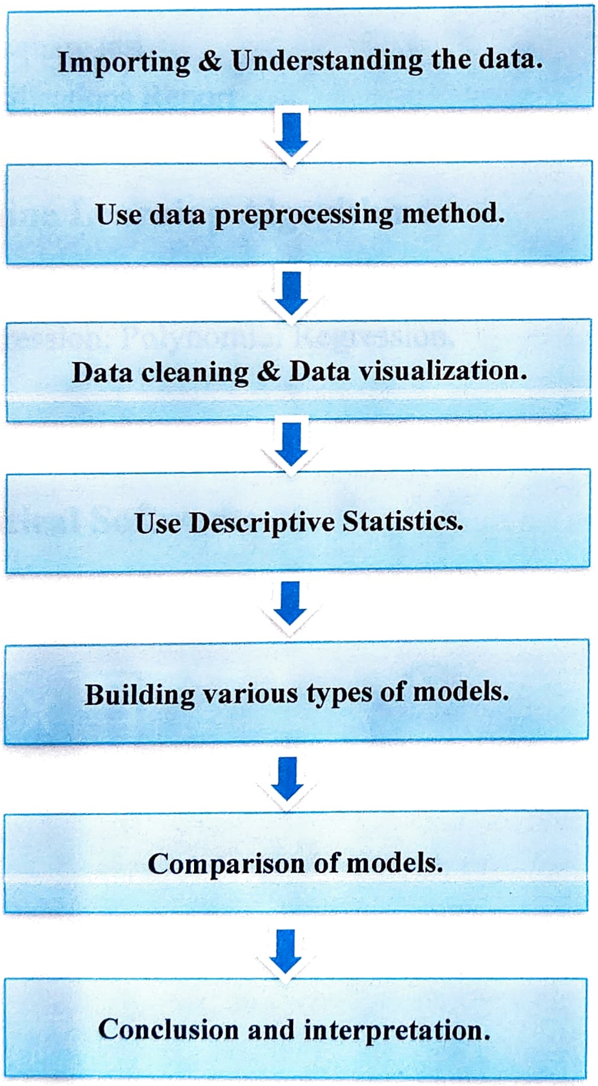
Website: www.kaggle.com

Variable Description:

- **Age:** The insured person's age.
- **Sex:** Gender (male or female) of the insured.
- **BMI (Body Mass Index):** A measure of body fat based on height and weight.
- **Children:** The number of dependents covered.
- **Smoker:** Whether the insured is a smoker (Yes or No).
- **Region:** The geographic area of coverage.
- **Charges:** The medical insurance costs incurred by the insured person.

METHODOLOGY

We will follow below methodology for this project Importing and understanding the data.



Importing & Understanding the data.

Use data preprocessing method.

Data cleaning & Data visualization.

Use Descriptive Statistics.

Building various types of models.

Comparison of models.

Conclusion and interpretation.

STATISTICAL TOOLS

Exploratory Data Analysis:

- Bar charts, Pie Charts, Correlation Heat map
- Chi-square test
- Classifications Report

Machine Learning Algorithms :

- Random Forest Regression, Linear Regression, Decision Tree Regression, Polynomial Regression.

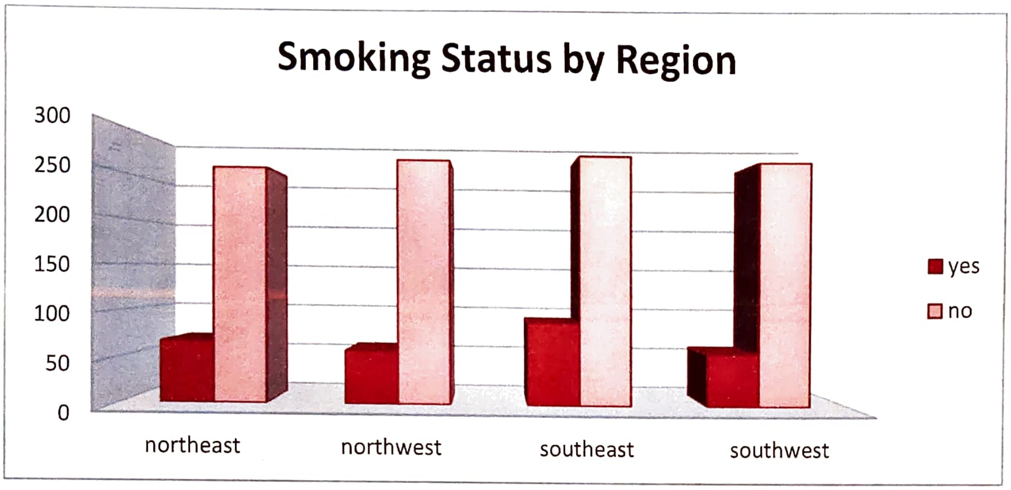
Statistical Software:

MS-Excel

Python

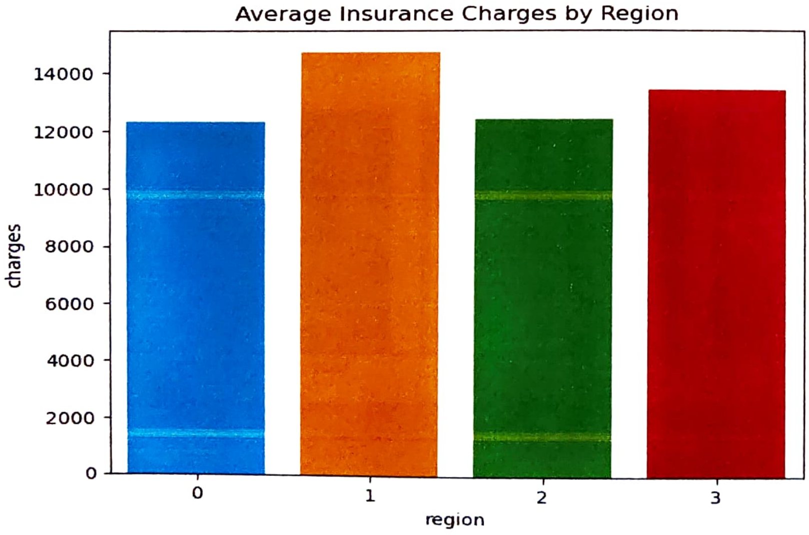
GRAPHICAL REPRESENTATION

A) Count of Smoking Status by Region:-



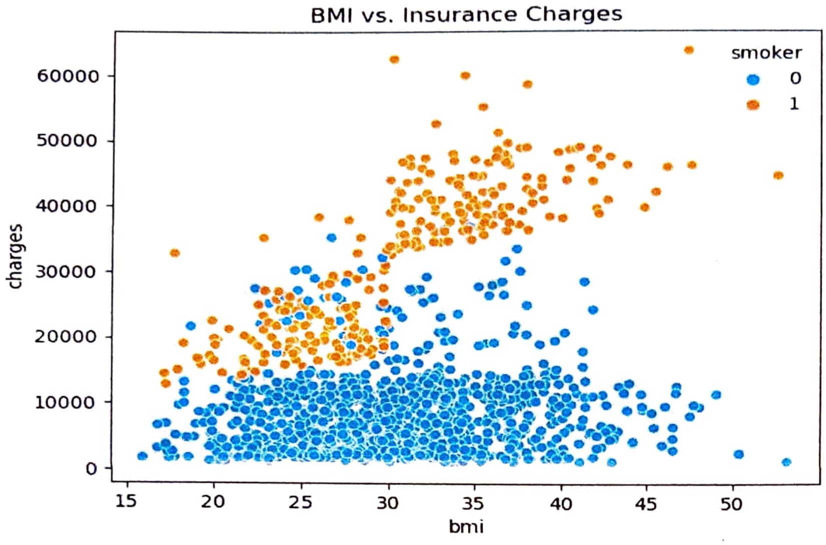
Conclusion: In all four regions, the number of non-smokers is significantly higher than the number of smokers.

B) Average Insurance Charges by Region:-



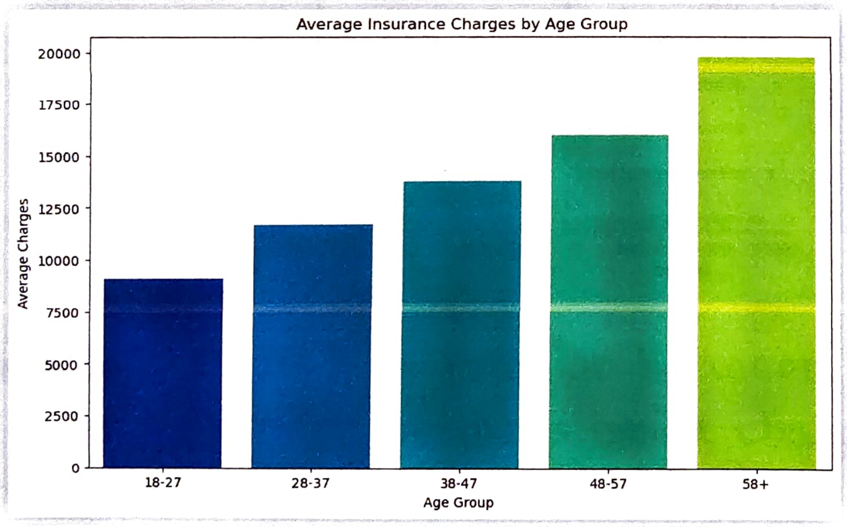
Conclusion: There is no huge difference between charges and region.

C) BMI vs. Insurance Charges:-



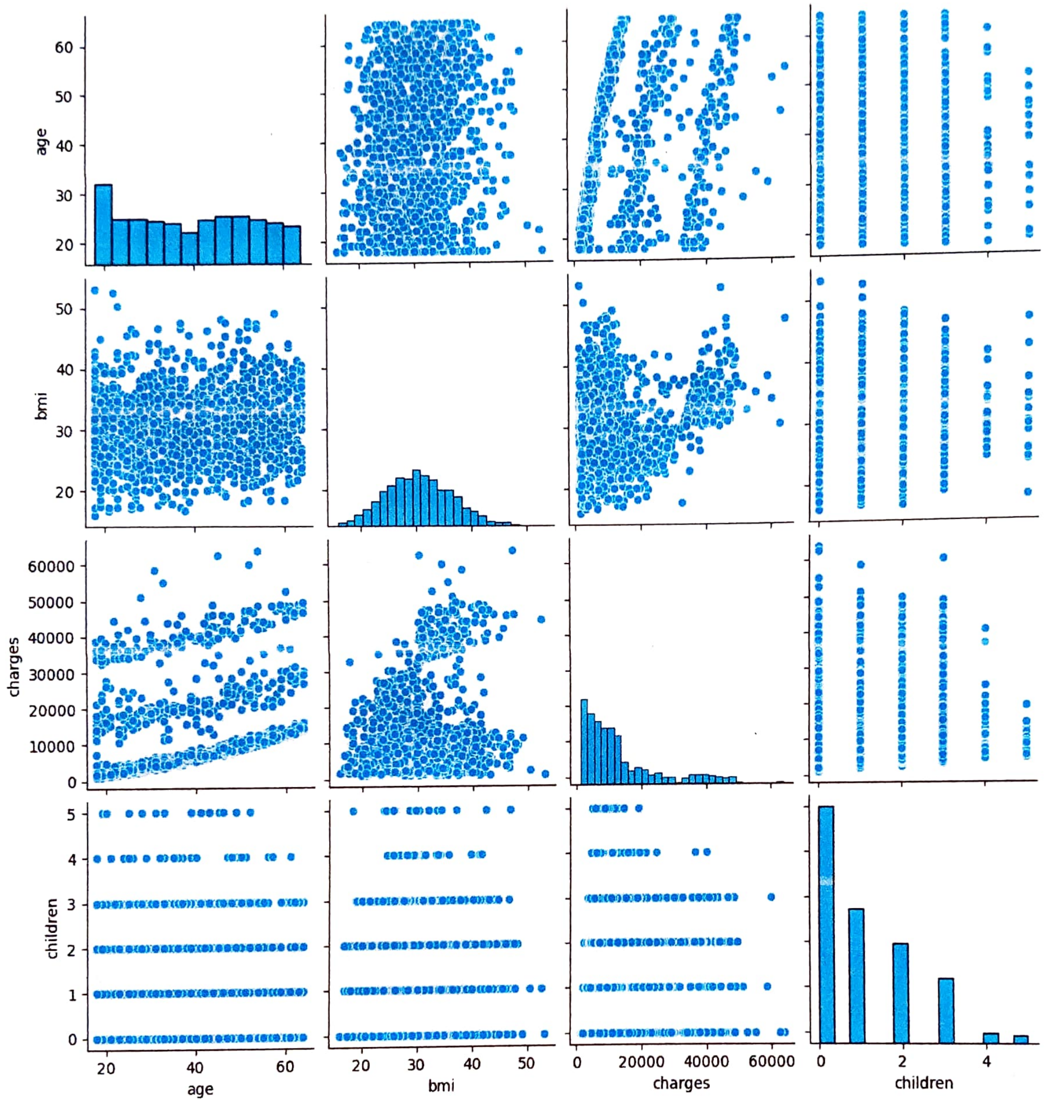
Conclusion: The scatter plot suggests a positive correlation between BMI and charges, with a general trend of increasing charges as BMI increases. Means the charges will increases if the BMI increases.

D) Average Insurance Charges by Age Group:-



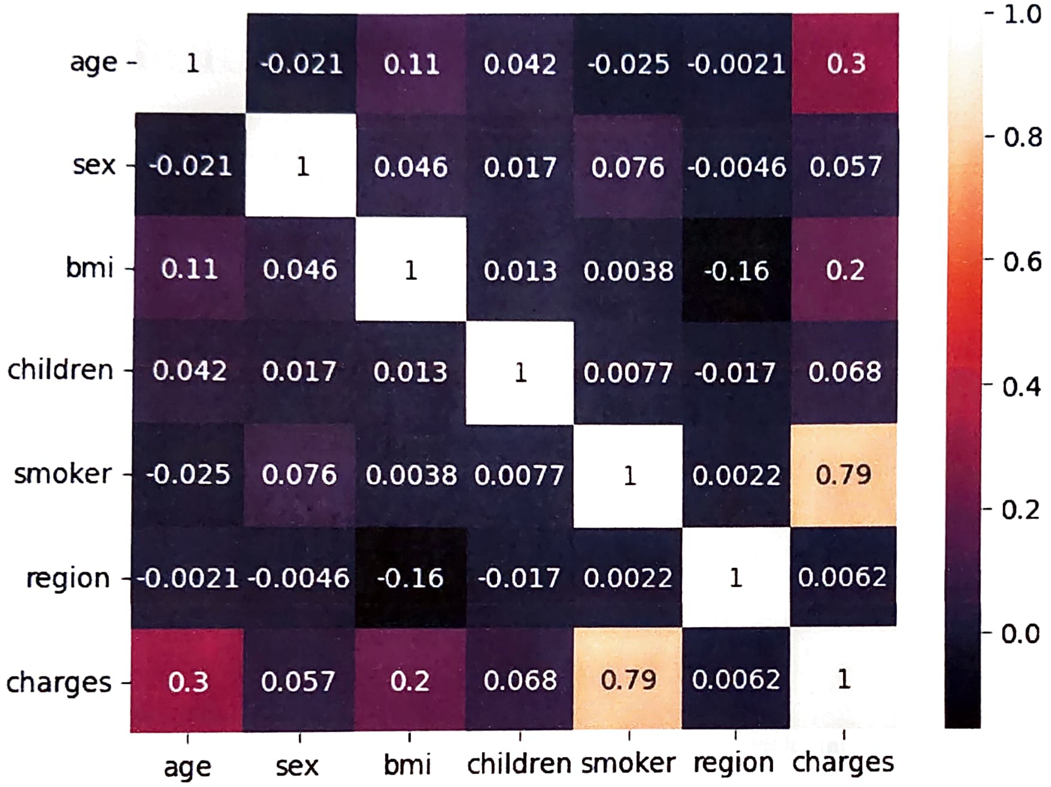
Conclusion: The average insurance charges generally increase with age. The group of 58-68 has highest insurance charges.

E) Paired Plot:-



Conclusion - Their is relation between age & bmi, and relation between bmi & charges.

CORRELATION HEATMAP:



Interpretation: The above map shows that their is correlation between Smoker & Charges.

STATISTICAL ANALYSIS

✦CHI-SQUARE TEST ANALYSIS

✓ **Testing Independency of Region & Smokers:-**

Hypothesis:

H0: Region & Smokers are independent on each other.

H1: Region & Smokers are dependent on each other.

Chi-square statistic: 7.3434

α: 0.05

P-value: 0.0671

P-value>0.05

Therefore accept H0 at 5% level of significance.

Conclusion: Region and Smokers are independent on each other.

Anova Test:-

Hypothesis:

Main Effect of Region:

H0: There is no significant difference in insurance charges across different regions.

H1: There is a significant difference in insurance charges across different regions.

F-statistic: 0.652722

α: 0.05

P-value: 0.5812827

P-value>0.05

Therefore accept H0 at 5% level of significance.

Conclusion: The p-value is greater than 0.05, so we fail to reject the H0. This suggests that there is no significant difference in insurance charges across different regions.

Main Effect of Smoker:

H0: There is no significant difference in insurance charges between smokers and non-smokers.

H1: There is a significant difference in insurance charges between smokers and non-smokers.

F-statistic: 2191.337326

p-value: 1.751867e-283

P-value<0.05

Conclusion: The p-value is much less than 0.05, so we reject the H0. This indicates a significant difference in insurance charges between smokers and non-smokers. Smokers have a significantly higher impact on insurance charges.

Interaction Effect of Region and Smoker:

H0: There is no interaction effect between region and smoker on insurance charges.

H1: There is an interaction effect between region and smoker on insurance charges.

F-statistic: 8.597631

P-value: 1.181560e-05

P-value<0.05

Conclusion: The p-value is less than 0.05, so we reject the H0. This suggests that there is a significant interaction effect between region and smoker status on insurance charges. This means that the effect of smoking on charges is different across regions.

Overall Conclusion:

- Region alone does not have a significant effect on insurance charges.
- Smoking status has a highly significant effect on insurance charges, with smokers incurring higher charges.
- There is a significant interaction effect between region and smoking status, indicating that the impact of smoking on insurance charges varies by region.

Effect of BMI & Smokers:

Conclusion:

- BMI significantly influences insurance charges. Different BMI levels result in varying charges.
- There is a significant interaction effect between BMI and smoker.
- The combined effect of BMI and smoker on insurance charges is significant, indicating that the effect of BMI on charges is different for smokers compared to non-smokers.

Tukey's Test for Region & Smokers:

Hypothesis:

H0 : There is no interaction effect between region and smoker status.

H1 : There is an interaction effect between region and smoker status.

Conclusion:

There is interaction effect between southeast and southwest.

❖ Multiple Linear Regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple linear regression refers to a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable.

Mathematical Imputation:

To improve prediction, more independent factors are combined. The following is the linear relationship between the dependent and independent variables.

$$y= \beta_{0+}\ \beta_1 x_1+\ \beta_2 x_2+\beta_3 x_3$$

- here, y is the dependent variable.
- x1, x2, x3… are independent variables.
- b0 =intercept of the line.
- b1, b2, … are coefficients.

Linear Regression Equation:

charges = 13402.8859 + (3492.3575 * age) + (-69.1754 * sex) + (2167.6623 * bmi) + (699.1658 * children) + (9738.9589 * smoker) + (343.5124 * region)

❖ Polynomial Regression:

Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables. The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x.

Polynomial Regression Equation:

charges = 236963404490217.5000 + (-144140864299.5693 * 1) + (3622.9421 * age) + (2107917830847.8601 * sex) + (1967.0899 * bmi) + (896.1798 * children) + (408186185655373.3125 * smoker) + (457.6962 * region) + (778.0552 * age^2) + (129.9557 * age sex) + (90.8775 * age bmi) + (-80.8043 * age children) + (15.5297 * age smoker) + (-315.7643 * age region) + (43361656064172.4844 * sex^2) + (70.3675 * sex bmi) + (-157.2145 * sex children) + (83.2732 * sex smoker) + (-127.8082 * sex region) + (-298.4218 * bmi^2) + (78.2122 * bmi children) + (3512.7386 * bmi smoker) + (244.4058 * bmi region) + (-136.3225 * children^2) + (-202.9520 * children smoker) + (295.2028 * children region) + (-280180919677181.1562 * smoker^2) + (-173.7178 * smoker region) + (131.7283 * region^2)

❖ Random Forest Regression:

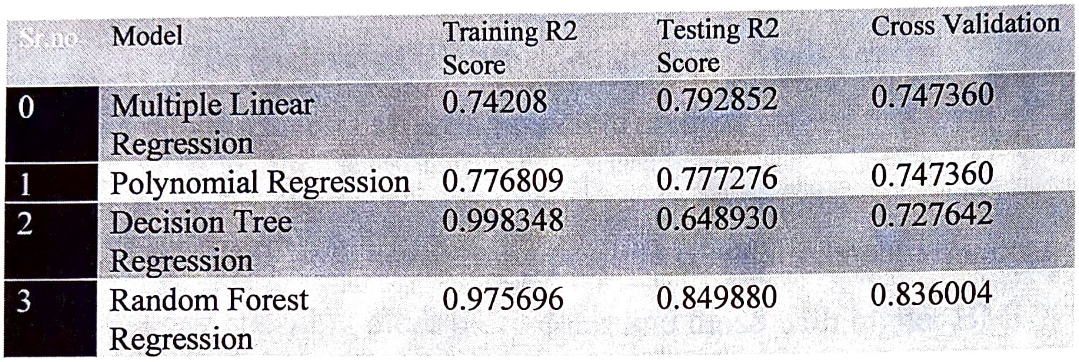
Random Forest Regression is an ensemble learning method for regression tasks that operates by constructing multiple decision trees during training. It outputs the mean prediction of the individual trees, which improves predictive accuracy and controls over-fitting.

Random Forest Regression is widely used in practice due to its robustness and ability to handle complex relationships in data.

❖ Decision Tree Regression:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. The goal is to create a tree structure that can accurately predict continuous outcomes based on input features.

✓ Regression Report:



Sr.no	Model	Training R2 Score	Testing R2 Score	Cross Validation
0	Multiple Linear Regression	0.74208	0.792852	0.747360
1	Polynomial Regression	0.776809	0.777276	0.747360
2	Decision Tree Regression	0.998348	0.648930	0.727642
3	Random Forest Regression	0.975696	0.849880	0.836004

- **Conclusion:** Among the models listed, Random forest seems to be the best choice as it maintains high performance on both training and testing data, with good generalized ability indicated by the cross-validation score.

CONCLUSION

The analysis of the provided dataset reveals several insights into the factors affecting medical charges

- ✓ Age & BMI Both have strong positive correlations with medical charges, meaning older individuals and those with higher BMI tend to have higher medical expenses.
- ✓ Smoking is a major factor that significantly increases medical charges.
- ✓ While age and BMI are strong influencers, the impact of sex.
- ✓ The interplay of these factors (age, BMI, smoker status, region) creates a complex picture of medical charges, underscoring the need for a comprehensive approach in understanding and predicting medical expenses.

REFERENCE

- K Swathi and R Anuradha (2017), Health insurance in India- An overview.
- Binny, Dr. Meenu Gupta (2017), Health insurance in India- Opportunities and challenges.
- BC Lakshmanna, P Jayarami Reddy, P Sravan Kumar (2019), Operational efficiency of Selected general insurance companies in India. Suman Devi and Dr. Vazir Singh
  Nehra (2015), The problems with health insurance sector in India.
- SatakshiChatterjee, Dr. ArunangshuGiri, Dr. S.N. Bandyopadhyay (2018), Health insurance sector in India: A study.