# Health Insurance Cost Prediction System Using Machine Learning

*A Project report submitted in the partial fulfilment of the requirements for the awardof the degree of*

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING**


Submitted by

**B. Kalyani (19471A0570)**

**K. Lalitha Annapurna(19471A0591)**

**A. Bhavani (19471A0568)**

Under the esteemed guidance of

**Dr.B.Jhansi Vazram** M.Tech.,Ph.D.

**Professor**




**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPET (AUTONOMOUS)**

Accredited by NAAC with A+ Grade and NBA under Cycle -1
NIRF rank in the band of 251-320 and an ISO 9001:2015 Certified
Approved by AICTE, New Delhi, Permanently Affiliated to JNTUK, Kakinada
KOTAPPAKONDA ROAD, YALAMANDA VILLAGE, NARASARAOPET-522601


2022-2023

**NARASARAOPET ENGINEERING COLLEGE: NARASARAOPET**
(AUTONOMOUS)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# CERTIFICATE


This is to certify that the project entitled "HEALTH INSURANCE COST PREDICTION SYSTEM USING MACHINE LEARNING" is a bonafide Work done by **"B. Kalyani (19471A0570), K. Lalitha Annapurna (19471A0591), A. Bhavani (19471A0568)"** in partial fulfilment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY** in the Department of **COMPUTER SCIENCE AND ENGINEERING** during the academic year 2022- 2023.

**PROJECT GUIDE**
Dr.B.Jhansi Vazram **M.Tech.,Ph.D.**
Professor

**PROJECT CO-ORDINATOR**
M.Sireesha **M.Tech., Ph.D.**
Associate Professor

**HEAD OF THE DEPARTMENT**
Dr. S. N.TirumalaRao **M.Tech,Ph.D.**

**EXTERNAL EXAMINER**

# ACKNOWLEDGEMENT

We wish to express our thanks to carious personalities who are responsible for the completion of the project. We are extremely thankful to our beloved chairperson sir **M.V.Koteswara Rao**, **B.sc** who took keen interest on us in every effort throughout this course. We owe out gratitude to our principal **Dr. Sreenivasa Kumar**, **M.Tech, Ph.D(UK),MISTE,FIE(1)** for his kind attention and valuable guidance throughout the course.

We express our deep felt gratitude to **Dr.S.N.Tirumala Rao**, M.Tech, Ph.D. H.O.D,CSE department and our guide **Dr.B.Jhansi Vazram** **M.Tech, Ph.D** of CSE department whose valuable guidance and unstinting encouragement enable us to accomplish our project successfully in time.

We extend our sincere thanks to **Ms.M.Sireesha** **M.Tech.,Ph.D.** Coordinator of the project for extending her encouragement. Their profound knowledge and willingness have been a constant source of inspiration for us throughout this project work.

We extend our sincere thanks to all other teaching and non-teaching staff of department for their cooperation and encouragement during our B.Tech degree. we have no words to acknowledge the warm affection, constant inspiration and encouragement that we receive from our parents.

We affectionately acknowledge the encouragement received from our friends and those who involved in giving valuable suggestions and clarifying out doubts, which had really helped us in successfully completing our project.

**By**
**B. Kalyani (19471A0570)**
**K. Lalitha Annapurna (19471A0591)**
**A. Bhavani (19471A0568)**

# ABSTRACT

The health care costs constitute a significant fraction of the U.S. economy. Nearly 20% of the Gross Domestic Product (GDP) is spent on health care. The health spending in theUS is the highest among all developed nations in absolute numbers as well as apercentage of the economy. In this work, we will develop a medical price prediction system using machine learning algorithms which will aid in steering patients to cost effective providers and thereby curb health spending. The policymakers can also use the tool to better understand which providers are relatively expensive and take punitive actions if necessary. The prediction of the medical price will be done using implementing Random Forest Regression algorithm in machine learning. Additionally, we plan to include the experiments on the same data with other machine learning models such as Random Forest Regression, Decision Tree Regression and Linear Regression and compare results. The findings from these experiments will also be included.

NARASARAOPETA
NEC ENGINEERING COLLEGE
(AUTONOMOUS)

# INSTITUTE VISION AND MISSION

## INSTITUTION VISION

To emerge as a Centre of excellence in technical education with a blend of effective student centric teaching learning practices as well as research for the transformation of lives and community.

## INSTITUTION MISSION

**M1**: Provide the best class infra-structure to explore the field of engineering and research.

**M2**: Build a passionate and a determined team of faculty with student centric teaching, imbibingexperiential, innovative skills.

**M3**: Imbibe lifelong learning skills, entrepreneurial skills and ethical values in students foraddressing societal problems.

NARASARAOPETA
NEC ENGINEERING COLLEGE
(AUTONOMOUS)

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

## VISION OF THE DEPARTMENT

To become a centre of excellence in nurturing the quality Computer Science & Engineering professionals embedded with software knowledge, aptitude for research andethical values to caterto the needs of industry and society.

## MISSION OF THE DEPARTMENT

The department of Computer Science and Engineering is committed to

**M1:** Mould the students to become Software Professionals, Researchers andEntrepreneurs byproviding advanced laboratories.

**M2:** Impart high quality professional training to get expertiZe in modern software tools andtechnologies to cater to the real time requirements of the industry.

**M3:** Inculcate team work and lifelong learning among students with a sense of societal and ethicalresponsibilities.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

# Program Specific Outcomes (PSO's)

**PSO1:** Apply mathematical and scientific skills in numerous areas of Computer Science andEngineering to design and develop software-based systems.

**PSO2:** Acquaint module knowledge on emerging trends of the modern era in ComputerScience and Engineering

**PSO3:** Promote novel applications that meet the needs of entrepreneur, environmental andsocial issues.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

# Program Educational Objectives (PEO's)

The graduates of the programme are able to:

**PEO1:** Apply the knowledge of Mathematics, Science and Engineering fundamentals toidentifyand solve Computer Science and Engineering problems.

**PEO2:** Use various software tools and technologies to solve problems related to academia,industry and society.

**PEO3:** Work with ethical and moral values in the multi-disciplinary teams and can communicate effectively among team members with continuous learning.

**PEO4:** Pursue higher studies and develop their career in software industry.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

# Program Outcomes

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis:** Identify, formulate, research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. **Design/development of solutions:** Design solutions for complex engineering problems anddesign system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems:** Use research-based knowledge and researchmethods including design of experiments, analysis and interpretation of data, and synthesis of theinformation to provide valid conclusions.

5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activitieswithan understanding of the limitations.

6. **The engineer and society:** Apply reasoning informed by the contextual knowledgeto assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to theprofessional engineering practice.

**7. Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and needfor sustainable development.

**8. Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

**9. Individual and team work**: Function effectively as an individual, and as a memberor leaderin diverse teams, and in multidisciplinary settings.

**10. Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

**11. Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

**12. Life-long learning**: Recognize the need for, and have the preparation and ability to engagein independent and life-long learning in the broadest context of technological change.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

## Project Course Outcomes (CO'S):

**CO425.1:** Analyse the System of Examinations and identify the problem.

**CO425.2:** Identify and classify the requirements.

**CO425.3:** Review the Related Literature**. CO425.4:** Design and Modularize the project.

**CO425.5:** Construct, Integrate, Test and Implement the Project.

**CO425.6:** Prepare the project Documentation and present the Report using appropriate method.

## Course Outcomes – Program Outcomes mapping

| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C425.1** | | ✓ | | | | | | | | | | | ✓ | | |
| **C425.2** | ✓ | | ✓ | | ✓ | | | | | | | | ✓ | | |
| **C425.3** | | | | ✓ | | ✓ | ✓ | ✓ | | | | | ✓ | | |
| **C425.4** | | | ✓ | | | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | |
| **C425.5** | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **C425.6** | | | | | | | | | ✓ | ✓ | ✓ | | ✓ | ✓ | |

## Course Outcomes – Program Outcome correlation

| | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C425.1** | 2 | 3 | | | | | | | | | | | 2 | | |
| **C425.2** | | | 2 | | 3 | | | | | | | | 2 | | |
| **C425.3** | | | | 2 | | 2 | 3 | 3 | | | | | 2 | | |
| **C425.4** | | | 2 | | | 1 | 1 | 2 | | | | | 3 | 2 | |
| **C425.5** | | | | | 3 | 3 | 3 | 2 | 3 | 2 | 2 | 1 | 3 | 2 | 1 |
| **C425.6** | | | | | | | | | 3 | 2 | 1 | | 2 | 3 | |

**Note: The values in the above table represent the level of correlation between CO's and PO's:**

1. Low level
2. Medium level
3. High level

## Project mapping with various courses of Curriculum with AttainedPO's:

| Name of the course from whichprinciples are applied in this project | Description of the device | Attained PO |
|---|---|---|
| C3.2.4, C3.2.5 | Gathering the requirements and definingthe problem, plan to develop a smart bottle for health care using sensors. | PO1, PO3 |
| CC4.2.5 | Each and every requirement is critically analyzed, the process model is identifiedand divided into five modules | PO2, PO3 |
| CC4.2.5 | Logical design is done by using theunified modelling language which involves individual team work | PO3, PO5, PO9 |
| CC4.2.5 | Each and every module is tested,integrated, and evaluated in our project | PO1, PO5 |
| CC4.2.5 | Documentation is done by all our four members inthe form of a group | PO10 |
| CC4.2.5 | Each and every phase of the work ingroup is presented periodically | PO10, PO11 |
| CC4.2.5 | Implementation is done and the projectwill be handled by the hospital management and in future updates in our project can be done based on airbubbles occurring in liquid in saline. | PO4, PO7 |
| CC4.2.8 CC4.2. | The physical design includes hardware components like sensors, gsm module,software and Arduino. | PO5, PO6 |

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# 1.INTRODUCTION

## 1.1 INTRODUCTION

We live on a planet full of threats and uncertainty. Including People, households, durables, properties are exposed to different risks and the risk levels can vary. These risks range from risk of health diseases to death if not get protection, and loss in property or assets. But risks cannot usually be avoided, so the world of finance has developed numerousproducts to shield individuals and organizations from these risks by using financial capitalto shield them. Therefore, Insurance is one of the policies that either decreases or removesloss costs incurred by various risks. The value of insurance in the lives of individuals. That'swhy it becomes important for insurance companies to be sufficiently precise to measure theamount covered by this specific policy and the insurance charges which must be paid for it.Various parameters or factors play an important role in estimating the insurance charges and Each of these is important. If any factor is omitted or changed when the amounts are computed then, the overall policy cost changes. It is therefore very critical to carry out thesetasks with high accuracy. So, the possibility of human mistakes is high so insurance agentsalso use different tools to calculate the insurance premium. And thus, ML is beneficial here.

ML may generalize the effort or method to formulate the policy. These ML models can be learned by themselves. The model is trained on insurance data from the past. The model can then accurately predict insurance policy costs by using the necessary elements to measure the payments as its inputs. This decreases human effort and resources and improves the company's profitability. Thus, the accuracy can be improved with ML. Our goal is to predict insurance costs. The value of insurance fees is based on different variables.As a result, insurance fees are continuous. Regression is the best choice available to fulfill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used.

Preprocessing of the dataset done first. Then we trained regression models with training data and finally evaluated these models based on testing data. In this article, we used several models of regression, for example, multiple linear regression, Decision Tree Regression and Gradient Boosting Regression and Classification models like Random Forest. It is found that the Random Forest provides the highest accuracy with an r-squared value of 86.7853. The inclusion of a novel method of insurance cost estimation is the main goal of this work.

## 1.2 EXISTING SYSTEM

Nowadays, Health insurance is emerging as a tool to manage financial needs of people to seek health services. Health Insurance is a part of "Personal Insurance and General Insurance". Decision is also made by themselves and this may lead to errors, and consume lot of time for getting the insurance for their treatment.

### Disadvantages

- Doesn't generate accurate and efficient results.
- Computation time is very high.
- Lacking of accuracy may result in lack of efficient further treatment

## 1.3 PROPOSED SYSTEM

By using this system, we can know their health status and their background details. By those details we can proceed to next step if their details are validated. If they are valid then the insurance will be given. If they are not valid then it will not proceed for giving insurance.

### Advantages:

- Generates accurate and efficient results.
- Computation time is greatly reduced.
- Reduces manual work.
- Efficient further treatment.

## 1.4 SYSTEM REQUIREMENTS

### 1.4.1 HARDWARE REQUIREMENTS

- Processor : Intel Core i5
- Cache Memory : 4MB
- Hard Disk : 30GB or more
- RAM : 1GB or more

### 1.4.2 SOFTWARE REQUIREMENTS

- Operating System : Window 10
- Coding Language : Python
- Python Distribution : Anaconda, Flask
- Browser : Any Latest Browser Like Chrome

# 2. LITERATURE SURVEY

## 2.1 MACHINE LEARNING

In this section, analysis efforts from the exploration of knowledge and Machine Learning techniques are mentioned. Many papers have discussed the difficulty of claim prediction. Jessica suggested, "Predicting motor insurance claims victimization telematics data". This research compared the performance of provision regression and XGBoost techniques to forecast the presence of accident claims by a little range and results showed that as a result of its interpretability and powerful predictability, logistic regression is a better model than XGBoost.

[4] System projected by Ranjodh Singh in 2019, this technique takes photos of the broken automobile as inputs and produces relevant details, akin to prices of repair, to come to a decision on the number of claims and locations of damage. so, the anticipated automobile insurance claim wasn't taken into consideration within the gift analysis however was focused on scheming repair costs.

Oskar Sucki 2019, the aim of this analysis is to check the prediction of churn. Random forests were thought-about to be the simplest model (75 % accuracy). In some fields, the information set had missing values. Following associate degree analysis of the distributions, the choice has been taken to substitute the missing variables with extra attributes suggesting that this data doesn't exist. This is often allowable given that the data is totally haphazardly way} lost, so the missing data mechanism by which the suitable approach to processing is set has 1st to be established.

In 2018, Muhammad rFauzan during this paper, the truth of XGBoost is applied to predict statements. Compare the output with the performance of XGBoost, a group of techniques e.g., AdaBoost, Random Forest, Neural Network. XGBoost offers higher Gini structured accuracy. mistreatment publically accessible urban centre Seguro to Kaggle datasets. The dataset includes vast quantities of NaN values however this paper manages missing values by medium and median replacement. However, these simple, unprincipled strategies have additionally proved to be biased. They, therefore, target exploring the cubic centimetre methods that are extremely applicable for the issues of many missing values, such as XGboost.

## 2.2 SOME MACHINE LEARNING METHODS

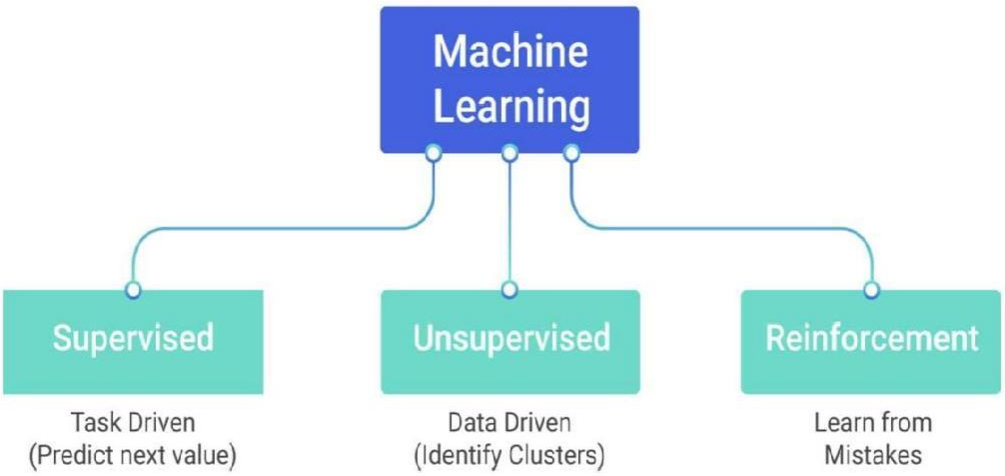


Figure: 2.1. Types of Machine Learning

Machine learning algorithms are often categorized as supervised and unsupervised.

- **Supervised machine learning algorithms:**

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

- **Unsupervised machine learning algorithms:**

Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function

to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

- **Reinforcement machine learning algorithms:**

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best. This is known as the reinforcement signal.

## 2.3 APPLICATIONS OF MACHINE LEARNING

a) Virtual Personal Assistants
b) Predictions while Commuting
c) Videos Surveillance
d) Social Media Services
e) Email Spam and Malware Filtering
f) Online Customer Support
g) Search Engine Result Refining
h) Product Recommendations
i) Online Fraud Detection

# 3. SYSTEM ANALYSIS

## 3.1 IMPORTANCE OF MACHINE LEARNING IN PYTHON

The importance of machine learning in healthcare is increasing because of its ability to process huge datasets efficiently beyond the range of human capability, and then dependably convert analysis of that data into clinical insights that assist physicians in planning and providing care, which ultimately leads to better outcomes, reduces the costs of care, and increases patients satisfaction. Using these types of advanced analytics, we can provide better information to doctors at the point of patient care.

## 3.2 IMPLEMENTATION OF MACHINE LEARNING USING PYTHON

Python is a popular programming language. It was created in 1991 by Guido van Rossum. It is used for:

1.web development (server-side), 2. software development, 3. mathematics,

4. system scripting.

The most recent major version of Python is Python 3. However, Python 2, although not being updated with anything other than security updates, is still quite popular.

It is possible to write Python in an Integrated Development Environment, such as Thonny, PyCharm, NetBeans or Eclipse, Anaconda which are particularly useful when managing larger collections of Python files.

Python was designed for its readability. Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.

Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

In the older days, people used to perform Machine Learning tasks manually by coding all

the algorithms and mathematical and statistical formula. This made the process time consuming, tedious and inefficient. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reasons is its vast collection of libraries. Python libraries that used in Machine Learning are:

1. Numpy 2. Scipy 3. Scikit-learn 4. Pandas 5. Matplotlib

**NumPy** is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

**SciPy** is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also very useful for image manipulation.

**Skikit-learn** is one of the most popular Machine Learning libraries for classical Machine Learning algorithms. It is built on top of two basic Python libraries, NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with Machine Learning.

**Pandas** is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides manyinbuilt methods for groping, combining and filtering data.

**Matpoltlib** is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, histogram, error charts, bar chats, etc.

## 3.3 SCOPE OF THE PROJECT

The scope of this system is to maintain patient details in datasets, train the model using the large quantity of data present in datasets and predict whether presence or absence of disease on new data during testing.

## 3.4 DATA SET ANALYSIS

We collected the data set from the Kaggle. Data Set consists of 7 variables. They are:

- **Age:** age of primary beneficiary.
- **Sex:** insurance contractor gender, female, male.
- **BMI:** Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9.
- **Children:** Number of children covered by health insurance/Number of dependents.
- **Smoker:** Is the person a smoker or not.
- **Region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **Charges:** Individual medical costs billed by health insurance.

## Data Set Link:

https://www.kaggle.com/code/shubhamptrivedi/health-insurance-price-predict-linear-regression

| age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|
| 19 | female | 27.9 | 0 | yes | southwest | 16884.92 |
| 18 | male | 33.77 | 1 | no | southeast | 1725.552 |
| 28 | male | 33 | 3 | no | southeast | 4449.462 |
| 33 | male | 22.705 | 0 | no | northwest | 21984.47 |
| 32 | male | 28.88 | 0 | no | northwest | 3866.855 |
| 31 | female | 25.74 | 0 | no | southeast | 3756.622 |
| 46 | female | 33.44 | 1 | no | southeast | 8240.59 |
| 37 | female | 27.74 | 3 | no | northwest | 7281.506 |
| 37 | male | 29.83 | 2 | no | northeast | 6406.411 |
| 60 | female | 25.84 | 0 | no | northwest | 28923.14 |
| 25 | male | 26.22 | 0 | no | northeast | 2721.321 |
| 62 | female | 26.29 | 0 | yes | southeast | 27808.73 |
| 23 | male | 34.4 | 0 | no | southwest | 1826.843 |
| 56 | female | 39.82 | 0 | no | southeast | 11090.72 |
| 27 | male | 42.13 | 0 | yes | southeast | 39611.76 |
| 19 | male | 24.6 | 1 | no | southwest | 1837.237 |
| 52 | female | 30.78 | 1 | no | northeast | 10797.34 |
| 23 | male | 23.845 | 0 | no | northeast | 2395.172 |
| 56 | male | 40.3 | 0 | no | southwest | 10602.39 |
| 30 | male | 35.3 | 0 | yes | southwest | 36837.47 |
| 60 | female | 36.005 | 0 | no | northeast | 13228.85 |
| 30 | female | 32.4 | 1 | no | southwest | 4149.736 |
| 18 | male | 34.1 | 0 | no | southeast | 1137.011 |
| 34 | female | 31.92 | 1 | yes | northeast | 37701.88 |
| 37 | male | 28.025 | 2 | no | northwest | 6203.902 |
| 59 | female | 27.72 | 3 | no | southeast | 14001.13 |
| 63 | female | 23.085 | 0 | no | northeast | 14451.84 |

Figure: 3.1 Data Set

## 3.5 METHADODLOGY

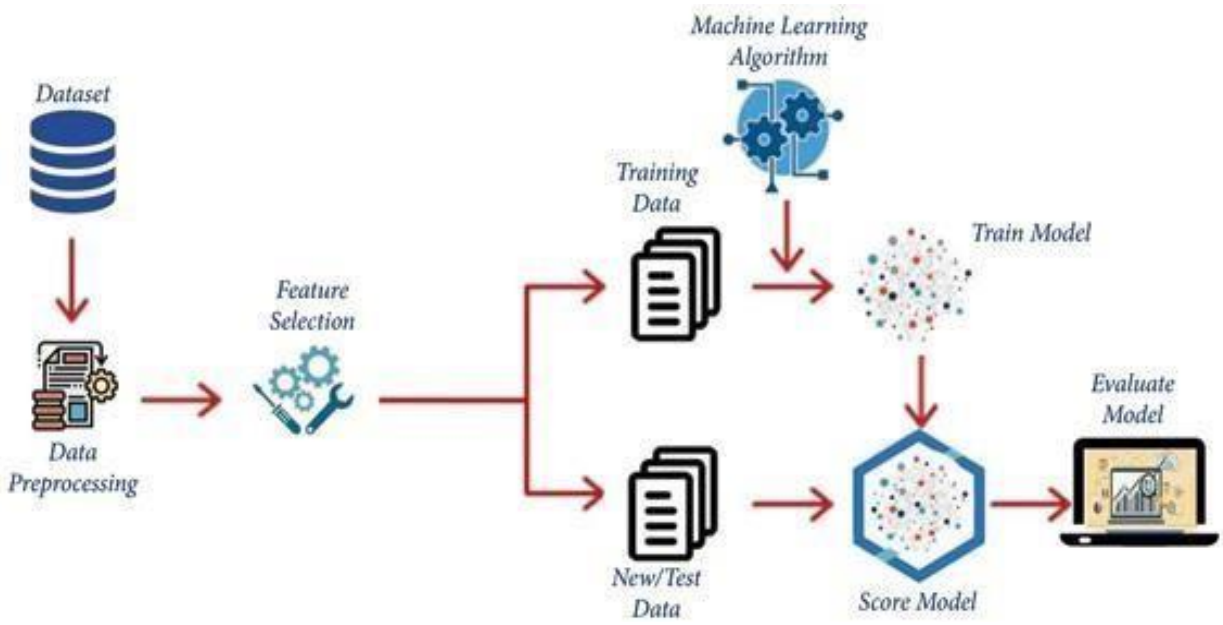


Figure: 3.2. Methodology

## 3.6 DATA PREPROCESSING

Before feeding data to an algorithm, we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing, the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.

| | |
|---|---|
| Age | Age of client |
| Sex | Male / Female<br>0=Male<br>1=Female |
| BMI | Body mass index |
| Children | Number of children the client have |
| Smoker | Whether or not a client smokest<br>0=yes<br>1=no |
| Region | Whether the client lives in southwest, northwest, southeast or northeast<br>0=southeast<br>1=southwest<br>2=northeast<br>3=northwest |
| Charges (Target Variable) | Medical Cost the client pay |

Table: 3.1. categorical variables after translated into numeric or binary values

## Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

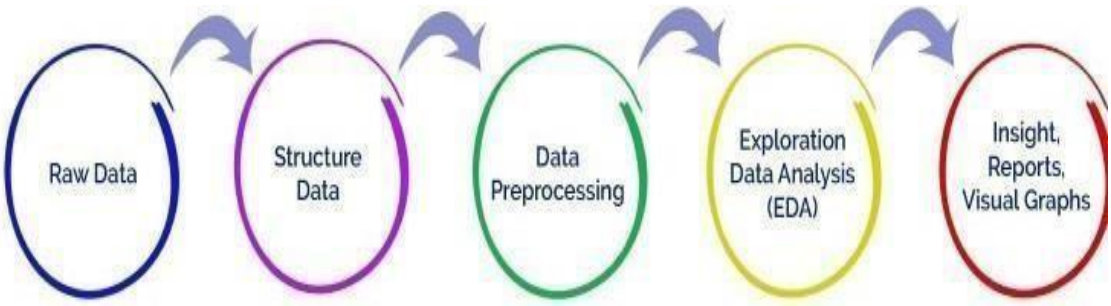


Figure: 3.3. Data Preprocessing

## Need of Data Preprocessing:

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format. For example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set. Another aspect isthat data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

## 3.7 CORRELATION COEFFICIENT METHOD

We can find dependency between two attributes p and q using Correlation coefficient method using the formula.

$rp, q= \sum(pi-p)(qi-q)/n\sigma p\sigma q$

$=\sum(pi\ qi)-np\ q/\ n\sigma p\sigma q$

n is the total number of patterns, pi and qi are respective values of p and q attributes in patterns i, p and q are respective mean values of p and q attributes, σp , σq are respective

standard deviations values of p and q attributes. Generally, -1≤ rp,q ≤ +1. If rp,q < 0, then p and q are negatively correlated. If rp,q =0, then p and q are independent attributes and there is no correlation between them. If rp,q > 0, then p and q are positively correlated. We can drop the attributes that are having correlation coefficient value as 0 as it indicates that the variables are independent with respect to the prediction attribute. Fig:3.8.2 is the correlation heat map. After applying correlation the attributes are PR interval , QRS duration , QT interval , QTc interval, P wave , T wave , QRS wave and problem . The attribute Vent_rate got dropped.


Figure: 3.4. Correlation

## 3.8 CROSS VALIDATION:

Cross-validation is a technique in which we train our model using the subset of the data- set and then evaluate using the complementary subset of the data-set. The three steps involved in cross-validation are as follows:

- Reserve some portion of sample data-set.
- Using the rest data-set train the model.
- Test the model using the reserve portion of the data-set.0

## 3.9 CLASSIFICATION

It is a process of categorizing data into given classes. Its primary goal is toidentify the class of our new data.

### Machine learning algorithms for classification

Research on data mining has led to the formulation of several data mining algorithms. These algorithms can be directly used on a dataset for creating some models orto draw vital conclusions and inferences from that dataset. Some popular data mining algorithms are Random Forest, Decision tree, Gaussian Naïve Bayes, Support vectormachine etc.

- **Decision Tree**

Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model.

- **Linear Regression Algorithm:**

Linear Regression is the first machine learning algorithm based on '**Supervised Learning'**.

Linear regression performs the task to predict a dependent variable value (y) based on a

given independent variable (x). Regression algorithms seem to be working on features represented as numbers only.

When there is a single input variable (x), the method is referred to as '**Simple Linear Regression'**. When there are multiple input variables, the method is referred to as **'Multiple Linear Regression'.**

**Step 1:** First we will split our data into 'X' array that contains the features and a 'y' array with the target variable.

**Step 2:** Next we will split our dataset(insurance.csv) into a training set and a testing set. We will train our model on the training set and then use the test set to evaluate the model (Predict 'y' variable). Please note that we will also compare the testing set predicted results with actual results. And import the required libraries.

**Step 3:** Train and Test the model

Now that we have a train and test datasets, we can evaluate the model using Linear regression

**Step 4:** Predictions from our Model

**Step 5:** Comparing the results

- ## Random Forest Regression:

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.

In this we use the Random Forest as Regression model. Random Forest's nonlinear nature can give it a leg up over linear algorithms, making it a great option. However, it is important to know your data and keep in mind that a Random Forest can't extrapolate.

Random Forest runs efficiently on large datasets. Random Forest has a high accuracy than other algorithms. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

## 3.10 TESTING DATA

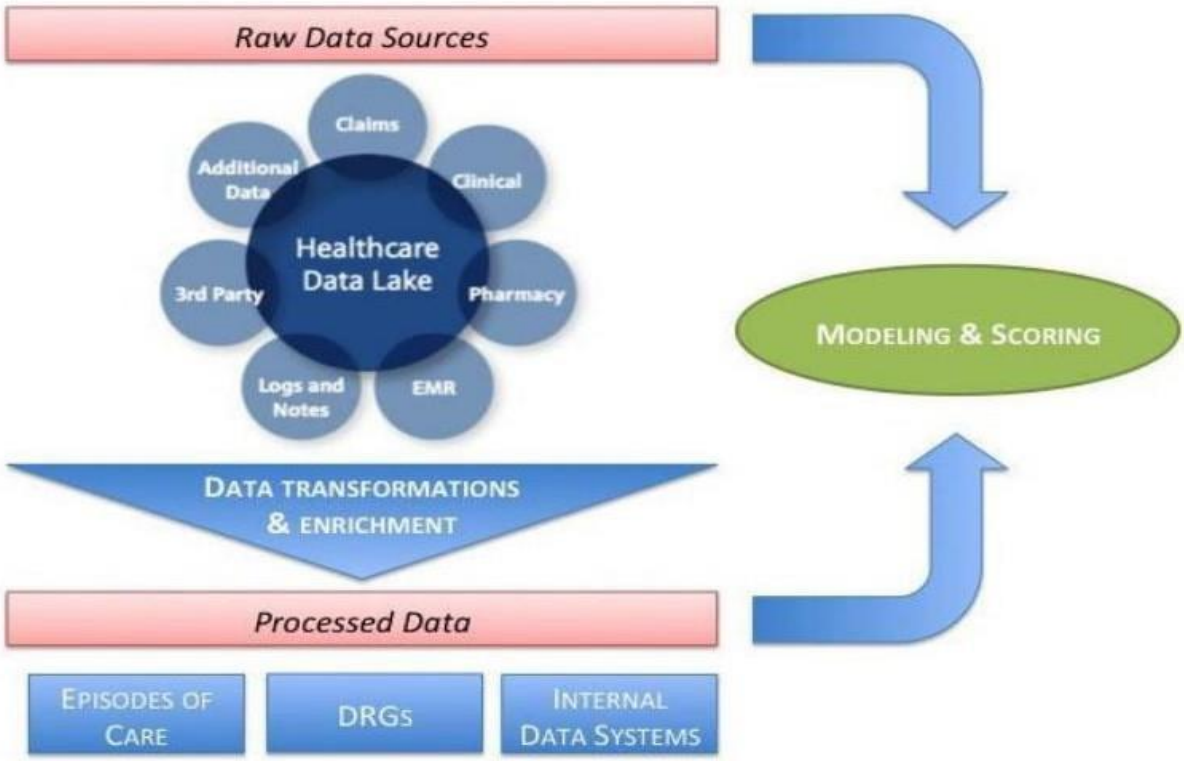


Figure: 3.5. Prediction model for medical insurance

Testing of data is done based on training model which is classified using supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in context of the given data.

## 3.11 IMPLEMENTATION

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib import style
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import KFold

df = pd.read_csv("insurance.csv")

df.head()

df.shape

df.columns

df.describe()

df.info()

plt.figure(figsize=(5,5))
style.use('ggplot')
sns.countplot(x='sex', data=df)
plt.title('Gender Distribution')
plt.show()

plt.figure(figsize=(5,5))
sns.countplot(x='smoker', data=df)
plt.title('Smoker')
plt.show()

plt.figure(figsize=(5,5))
```

```
sns.countplot(x='region', data=df)
plt.title('Region')
plt.show()

plt.figure(figsize=(5,5))
sns.barplot(x='sex', y='charges',hue='smoker', data=df)
plt.title('Charges for smokers')

df[['age','bmi','children','charges','sex','smoker','region']].hist(bins=30, figsize=(10,10),
color='blue')
plt.show()

df.head()

df['sex'] = df['sex'].apply({'male':0, 'female':1}.get)
df['smoker'] = df['smoker'].apply({'yes':1, 'no':0}.get)
df['region'] = df['region'].apply({'southwest':1, 'southeast':2, 'northwest':3,
'northeast':4}.get)

plt.figure(figsize=(10,7))
sns.heatmap(df.corr(), annot = True)
plt.show()

df.plot(kind="box",subplots=True,sharex=False,sharey=False,figsize=(20,10),color='deep
pink')

X = df.drop(['charges'], axis=1)
y = df.charges

x_train, x_test, y_train, y_test = train_test_split(X,y, test_size=0.3, random_state=42)
print("X_train shape: ", x_train.shape)
print("X_test shape: ", x_test.shape)
print("y_train shape: ", y_train.shape)
print("y_test shape: ", y_test.shape)

linreg = LinearRegression()

linreg.fit(x_train, y_train)
x_pred=linreg.predict(x_train)
y_pred = linreg.predict(x_test)
```

```
print("train score:",linreg.score(x_train,y_train))
print("test score:",linreg.score(x_test,y_test))

print('MAE= ',metrics.mean_absolute_error(y_test,y_pred))
print('MSE= ',metrics.mean_squared_error(y_test,y_pred))
print(f"r2 score: {r2_score(y_test,y_pred)}")
print('Adjusted R2 value= ',1 - (1 - (linreg.score(x_test,y_test))) * ((756 - 1)/(756-10-1)))
print('RMSE (train)= ',np.sqrt(mean_squared_error(y_train,x_pred)))
print('RMSE (test)= ',np.sqrt(mean_squared_error(y_test,y_pred)))

data = {'age':50,'sex':0, 'bmi':25, 'children':2, 'smoker':1, 'region':2}
index = [0]
cust_df = pd.DataFrame(data, index)
cust_df

cost_pred = linreg.predict(cust_df)
print("The medical insurance cost of the new customer is: ", cost_pred)

# Fitting Random Forest Regression to the dataset
regressor = RandomForestRegressor(n_estimators = 100)
regressor.fit(x_train,y_train)
x_pred=regressor.predict(x_train)
y_pred=regressor.predict(x_test)
print("train score:",regressor.score(x_train,y_train))
print("test score:",regressor.score(x_test,y_test))

print('MAE= ',metrics.mean_absolute_error(y_test,y_pred))
print('MSE= ',metrics.mean_squared_error(y_test,y_pred))
print(f"r2 score: {r2_score(y_test,y_pred)}")
print('Adjusted R2 value= ',1 - (1 - (regressor.score(x_test,y_test))) * ((756 - 1)/(756-10-1)))
print('RMSE (train)= ',np.sqrt(mean_squared_error(y_train,x_pred)))
print('RMSE (test)= ',np.sqrt(mean_squared_error(y_test,y_pred)))

data = {'age':50,'sex':0, 'bmi':25, 'children':2, 'smoker':1, 'region':2}
index = [0]
cust_df = pd.DataFrame(data, index)
cust_df
```

```
cost_pred = regressor.predict(cust_df)
print("The medical insurance cost of the new customer is: ", cost_pred[0])


# Fitting Decision Tree Regression to the dataset
regressordt = DecisionTreeRegressor()
regressordt.fit(x_train,y_train)
x_pred=regressordt.predict(x_train)
y_pred=regressordt.predict(x_test)
print("train score:",regressordt.score(x_train,y_train))
print("test score:",regressordt.score(x_test,y_test))

print('MAE= ',metrics.mean_absolute_error(y_test,y_pred))
print('MSE= ',metrics.mean_squared_error(y_test,y_pred))
print(f"r2 score: {r2_score(y_test,y_pred)}")
print('Adjusted R2 value= ',1 - (1 - (regressordt.score(x_test,y_test))) * ((756 - 1)/(756-10-1)))
print('RMSE (train)= ',np.sqrt(mean_squared_error(y_train,x_pred)))
print('RMSE (test)= ',np.sqrt(mean_squared_error(y_test,y_pred)))

kf=KFold(n_splits=7)
kf

for train_index,test_index in kf.split(['age','sex','bmi','children','smoker','region','charges']):
    print(train_index,test_index)

def get_score(model,x_train,x_test,y_train,y_test):
    model.fit(x_train,y_train)
    return model.score(x_test,y_test)

get_score(LinearRegression(),x_train,x_test,y_train,y_test)

get_score(RandomForestRegressor(n_estimators = 100),x_train,x_test,y_train,y_test)

get_score(DecisionTreeRegressor(),x_train,x_test,y_train,y_test)
```

## app.py

```
from flask import Flask, request, render_template
import pickle

app = Flask(_name_, template_folder='template')

model = pickle.load(open("model.pkl", "rb"))


@app.route('/')
def home():
    return render_template('index.html')


@app.route('/health')
def health():
    return render_template('health.html')


@app.route('/predict', methods = ["POST", "GET"])
def predict():
    age = int(request.form['age'])
    gender = int(request.form['gender'])
    bmi = float(request.form['bmi'])
    children = int(request.form['children'])
    smoke = int(request.form['smoke'])
    region = int(request.form['region'])
    result = model.predict([[age, gender, bmi, children, smoke, region]])
    return render_template('submit.html',  result="$ {:.2f}".format(result[0]))


if _name_ == '_main_':
    app.run(debug=True)
```

## index.html

```
<!DOCTYPE html>
<html lang="en" xmlns:https="http://www.w3.org/1999/xhtml">
<head>
    <meta charset="UTF-8">
    <title>Life Line</title>
    <link href='https://fonts.googleapis.com/css?family=Josefin Sans' rel='stylesheet'>
    <link rel="stylesheet" href="{{ url_for('static',filename='css/style.css') }}">
</head>

<body>
```

```
<nav>
   <label class="logo">LIFE LINE</label>
</nav>
<section class="about">
   <div class="main">
      <img src="https://images.moneycontrol.com/static-mcnews/2021/11/Health-
insurance-770x433.jpg?
      impolicy=website&width=770&height=431">
      <div class="about-text">
         <h1>About Us</h1>
         <h5>Health Insurance <span>Cost Prediction</span></h5>
         <p>Life Line is a health insurance cost prediction system. This website helps you
to predict health insurance cost of the user
         with the help of the attributes given by the user. To know your health insurance
cost click on get started.</p>
         <button type="button" onclick="window.location.href='{{ url_for('health')
}}';">Get Started</button>
      </div>
   </div>
</section>

</body>
</html>
```

## style.css

```
*{
   padding: 0;
   margin: 0;
   font-family: 'Josefin Sans';
   box-sizing: border-box;


}
.about{
   width: 100%;
   padding: 100px 0px;
   background-color: #1919191;
}
.about img{
   height: 300px;
   width: 500px;
}

.about-text{
   width: 550px;
}
.main{
   width: 1130px;
```

```
    max-width: 95%;
    margin: 0 auto;
    display: flex;
    align-items: center;
    justify-content: space-around;
}
.about-text h1{
    color: black;
    font-size: 80px;
    text-transform: capitalize;
    margin-bottom: 20px;
}
.about-text h5{
    color: black;
    font-size: 25px;
    text-transform: capitalize;
    margin-bottom: 25px;
    letter-spacing: 2px;
}
span{
    color: #ff5c5c;
}
.about-text p{
    color: black;
    letter-spacing: 1px;
    line-height: 28px;
    font-size: 18px;
    margin-bottom: 45px;
}
button{
    background: #ff5c5c;
    color: black;
    text-decoration:  none;
    border: 2px solid transparent;
    font-weight: bold;
    padding: 13px 30px;
    border-radius:  30px;
    transition: .4s;
}
button:hover{
    background: transparent;
    border: 2px solid #ff5c5c;
    cursor: pointer;
}
nav{
    font-family: 'Bradley Hand, cursive';
    height: 70px;
    width: 100%
}
label.logo{
```

```
    color: #ff5c5c;
    font-size: 35px;
    line-height: 80px;
    padding-left:30px;
    font-weight: bold;
}
```

## health.html

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Life Line</title>
    <link href='https://fonts.googleapis.com/css?family=Josefin Sans' rel='stylesheet'>
    <link rel="stylesheet" href="{{ url_for('static',filename='css/form_style.css') }}">
    <nav>
    <label class="logo">LIFE LINE</label>
</nav>
</head>

<body>
<div class="heading">
    <h1>Predict the cost of your Medical Insurance!</h1>
</div>
<form action='/predict' method="POST">
    <div class="parent">
     <div class="child">
    <label>Age :</label><br><br>
    <input type="number" name="age" placeholder="      Age > 18" min="18"
required="required" class="input"><br>
    </div>
    <div class="child" >
    <label>Gender :</label><br><br>
    <input type="number" name="gender" placeholder="      0-Male / 1-Female" min="0"
max="1" required="required" class="input" ><br>
    </div>
    <div class="child">
    <label>BMI :</label><br><br>
    <input type="number" name="bmi" placeholder="      BMI" required="required"
class="input"><br>
     </div>
    </div>
    <div class="parent2">
     <div class="child2">
    <label>Children :</label><br><br>
    <input type="number" name="children" placeholder="       0 for None"
required="required" class="input"><br>
```

```
    </div>
    <div class="child2">
    <label>Do you smoke ?</label><br><br>
    <input type="number" name="smoke" placeholder=" 1-Yes / 0-No" max="1"
min="0" required="required" class="input"><br>
    </div>
        <div class="child2">
    <label>Region :</label><br><br>
    <input type="number" name="region" placeholder=" 1-SW / 2-SE / 3-NW / 4-NE"
max="4" min="1" required="required" class="input"><br>
    </div>
    </div>
    <div class="bt">
    <button type="submit">Predict</button>
    </div>
</form>
</body>
</html>
```

## form_style.css

```
body{
    background-image:
url('https://i.pinimg.com/736x/23/e8/20/23e820216a84f32bc11077c20c0e3f0e.jpg');
    background-repeat:  no-repeat;
    background-attachment: fixed;
    background-size: cover;
}
*{
    padding: 0;
    margin: 0;
    font-family: 'Josefin Sans';
    box-sizing: border-box;
}
nav{
    font-family: 'Bradley Hand, cursive';
    height: 70px;
    width: 100%
}
label.logo{

    color: #03254c;
    font-size: 35px;
    line-height: 80px;
    padding-left:30px;
    font-weight: bold;
}
```

```
.heading
{
   text-align: center;
   padding-top: 40px;
   padding-right: 30px;
   color: #67032f  ;
}
button{
   background: #67032f;
   color: white;
   text-decoration:  none;
   border: 2px solid transparent;
   font-weight: bold;
   padding: 13px 30px;
   border-radius:  30px;
   transition: .4s;

}
button:hover{
   background: transparent;
   border: 2px solid #67032f;
   cursor: pointer;
}
form{
   padding-left:100px;
   padding-top: 20px;
   font-size: 20px;
}
.input{
   height: 30px;
   width:220px;
   font-size:15px;

}
.parent{

   margin: 1rem;
   padding: 2rem 2rem;
   text-align: center;
   padding-right:170px;
}
.child{
   display:inline-block;

   padding: 1rem 2rem;
   vertical-align: middle;
   padding-right: 40px;
}

.parent2{
```

```
    margin: 1rem;
    padding: 2rem 2rem;
    text-align: center;
    padding-right:170px;

}
.child2{
    display:inline-block;

    padding: 1rem 2rem;
    vertical-align: middle;
    padding-right: 40px;
}
.bt{
    padding-left:460px;
    padding-top:40px;
}
```

## submit.html

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Life Line</title>
    <link href='https://fonts.googleapis.com/css?family=Josefin Sans' rel='stylesheet'>

    <nav>
    <label class="logo">LIFE LINE</label>
</nav>
</head>
<style>
    body{
    background-image: url('https://www.statnews.com/wp-
content/uploads/2021/04/AdobeStock_340574222-1-1600x900.jpeg');
    background-repeat:  no-repeat;
    background-attachment: fixed;
    background-size: cover;
}
*{
    padding: 0;
    margin: 0;
    font-family: 'Josefin Sans';
    box-sizing: border-box;
}
nav{
```

```
    font-family: 'Bradley Hand, cursive';
    height: 70px;
    width: 100%
}
label.logo{

    color: #990000;
    font-size: 35px;
    line-height: 80px;
    padding-left:30px;
    font-weight: bold;
}
h2{
    padding-top:50px;
    padding-left:440px;
}
h1{
    padding-top:60px;
    padding-left:540px;
    color: #990000;
}
</style>
<body>
    <h2>Predicted Health Insurance Cost is</h2>
    <h1>{{result}}</h1>
</body>
</html>
```

## 3.12 RESULT ANALYSIS

The performance of the several regressors were analysed in order to pick the most effective fundamental algorithm and establish the correctness and reliability of the model's deployment. The efficiency of the regressor can be explained using the estimation coefficient and the Root Mean Squared Error (RMSE) (r2 score).

| MODEL | R2 SCORE | RMSE | ACCURACY |
|---|---|---|---|
| Linear Regression | 0.7694415 | 6144.199 | 76% |
| Random Forest Regression | 0.8525963 | 1876.602 | 85 % |
| Decision Tree Regression | 0.7293854 | 6299.100 | 72% |

Table: 3.2. Comparison of models

Comparing the multiple regression models used in Table 2 for predicting the insurance claim revealed that the larger the r-squared value of the models. Hence, the R-squared number gives the most accurate representation of how variable the dependent variables is in relation to the surrounding mean. With an R2 value of 0.8525963, the Random Forest Regression model was determined to be the best model.

# 4. OUTPUT SCREENS

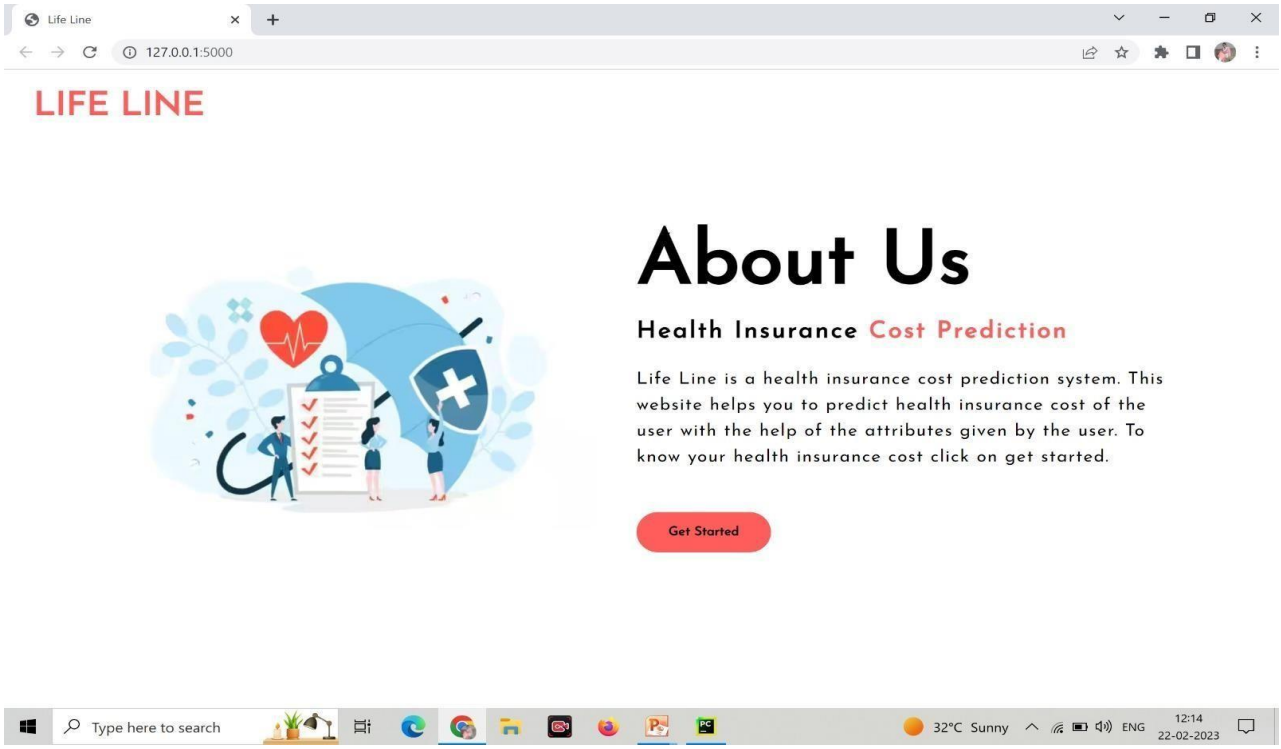


Figure: 4.1. Home Screen

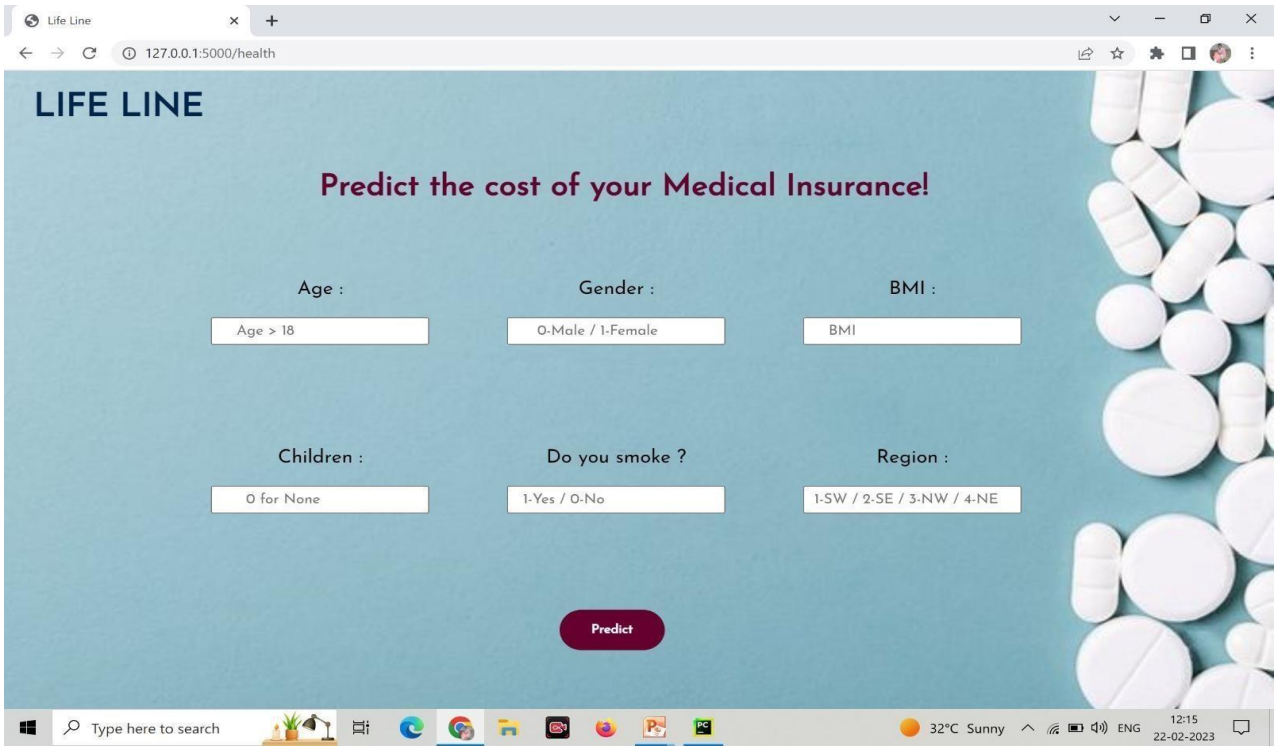


Figure: 4.2. Prediction Form

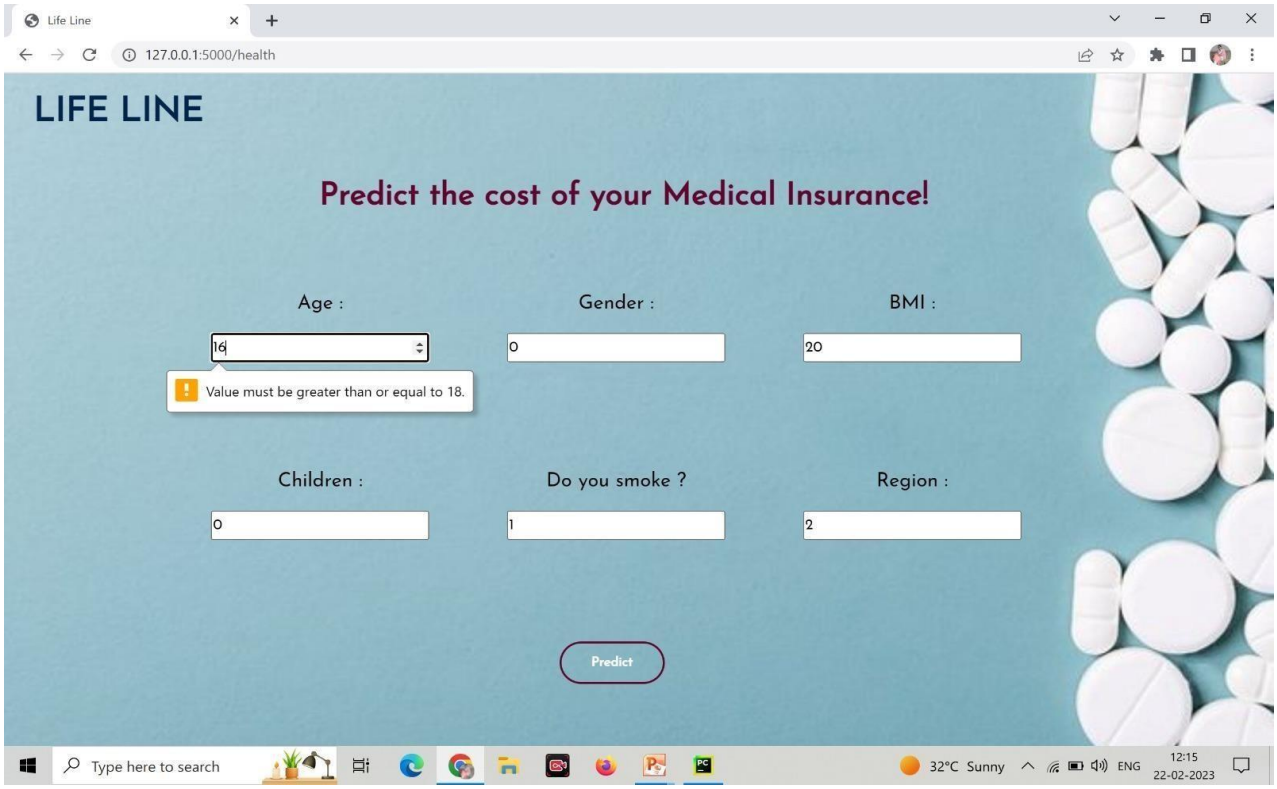


Figure: 4.3. Form Validation

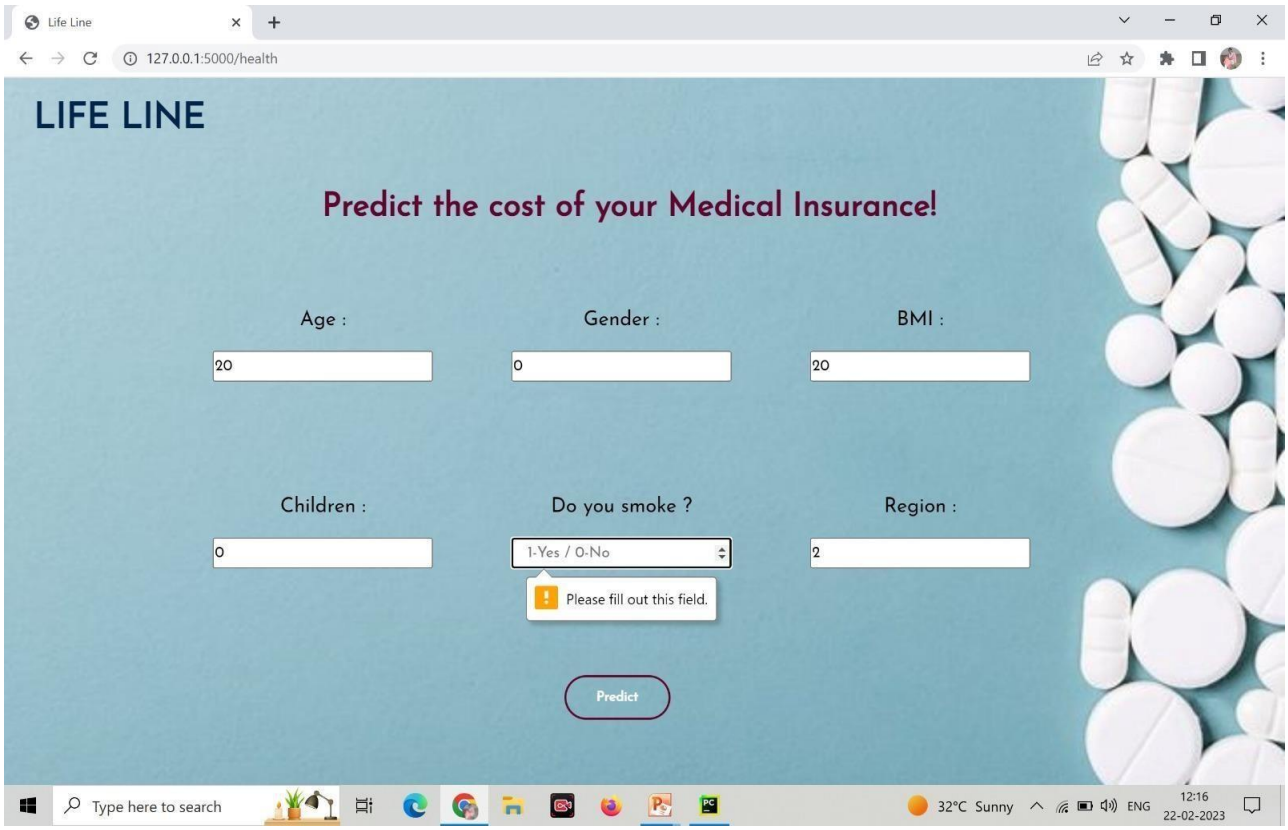


Figure: 4.4. Missing Field Validation

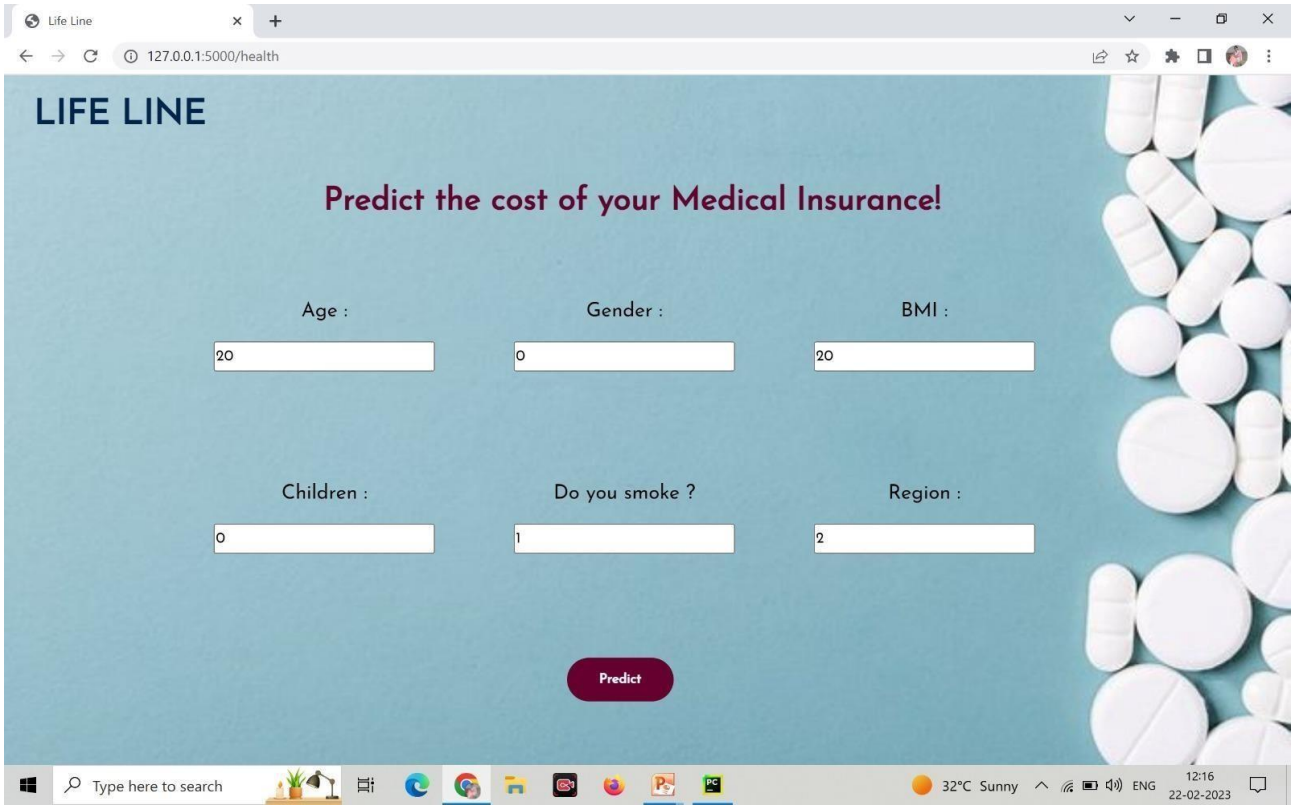


Figure: 4.5. Inserting values in the form

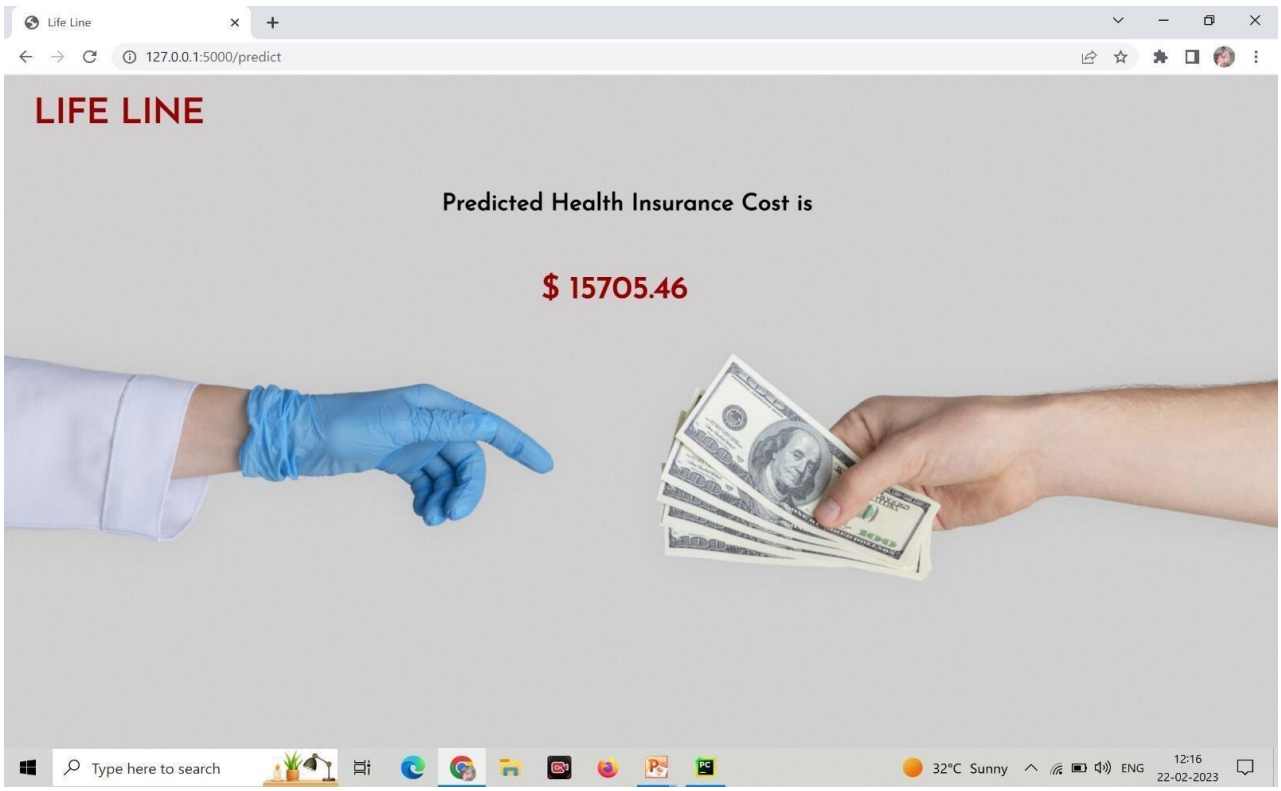


Figure: 4.6. Output Screen

# 5.CONCLUSION AND FUTURE SCOPE

## 5.1 Conclusion

The expense of the health policy has been estimated as precisely as possible using the predictive models described here. To create better medical facilities, this is allegedly highly beneficial for the healthcare organisation. The study made use of Decision Tree Regression, Random Forest Regression, and Linear Regression methodologies. When all the results were compared, Random Forest Regressor had the greatest R2 score.

## 5.2 Future Scope

Future research, however, can focus on exploring a number of additional deep learning applications. The technique that can converge the changes and do multiple framework can be improved using improved approaches from machine learning and other fields. Although most forecast errors involve a significant financial claim transaction, updating some of them can generally benefit insurance firms.

# 6.BIBLIOGRAPHY


1. https://www.kaggle.com/code/shubhamptrivedi/health-insurance-price-predict-linear-regression

2. "Prediction of Insurance Claim Severity Loss Using Regression Models," R. M. Ogunnaike and D. Si, 2017, pp. 233-247.

3. "Modeling frequency and severity of claims with the zero-inflated generalised cluster weighted models," Insur. Math. Econ., vol. 94, pp. 79–93, September 2020; doi: 10.1016/j.insmatheco.2020.06.004..

4. "A complete overview and analysis of generative models in machine learning," Computer Science Review, vol. 38, no. 100285, 2020, doi: 10.1016/j.cosrev.2020.100285.

5. "Nuclei segmentation in cell pictures using fully convolutional neural networks," Int. J. Emerg. Technol., vol. 11, no. 3, pp. 731–737, 2020. S. S. Rautaray, S. Dey, M. Pandey, and M. K. Gourisaria.

6. Evaluation of Technology Adoption Models and Theories to Assess Readiness and Appropriate Usage of Technology in a Corporate Organization by T. Dube, R. Van Eck, and T. Zuva 10.36548/jitdw.2020.4.003

7. T. J. Layton, "Imperfect risk adjustment, risk preferences, and sorting in competitive health insurance markets," J. Inf. Technol. Digit. World, vol. 02, no. 4, pp. 207–212, 2020. 56, 259–280, J. Health Econ., 2017, doi: 10.1016/j.jhealeco.2017.04.004.

8. Jayasree, M., and M. G. Chandrasekhar (2020). Machine Learning Methods for Predicting Health Insurance Claims. The International Conference on Communication and Signal Processing (ICCSP) will take place in 2020. (pp. 0426-0431). IEEE.

9. P. Awasthi, P. Sharma, and S. (2020). Machine Learning for Predictive Analysis of Health Insurance Claims. The eleventh ICCCNT (International Conference on Computation, Communication, and Networking Technologies) will take place in 2020. (pp. 1-6). IEEE

10. Albarrak, A. M., Elshaikh, E. A., and Basheer, M. A. (2021). A thorough review of the prediction of health insurance claims using machine learning techniques. 45(1), 1-14 in Journal of Medical Systems.

# *Health Insurance Claims Prediction Using Machine Learning*

**B.Kalyani[1],K.Lalitha Annapurna[2],A.Bhavani[3],B.Jhansi Vazram[4]**

[1,2,3]Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

[4]Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

bandikalyani392@gmail.com[1] ,klalithaannapurna@gmail.com[2],ambatibhavani1111@gmail.com[3], jhansi.bolla@gmail.com[4]

**1. ABSTRACT-** The goal of this project is to create a predictive model for health insurance claims using machine learning methods. The Kaggle website provided the dataset for this study. The technology can also help policymakers identify which providers are often more expensive and, if required, take punitive action. To produce useful features for our machine learning models, we preprocess the data and engage in feature engineering. Following that, we assess a number of regression techniques utilizing metrics, including linear regression, random forest regression, and decision tree regression.

**KEYWORDS**: Health Insurance Claims, Machine Learning, Linear Regression, Random forest regression,Decision tree regression, Probability Prediction

## 2. INTRODUCTION

Machine learning, a fast evolving field of artificial intelligence, enables computers to automatically acquire knowledge through experience and improve over time without human input. Machine learning allows computers to examine massive volumes ofdata, spot patterns and trends, and then use that information to predict the future or make decisions. There are many useful uses for machine learning, including fraud detection, natural language processing, picture and speech recognition, and personalised recommendations. Machine learning is anticipated to have a significant impact on many businesses and facets of daily life as it develops.

A healthcare industry use of artificial intelligence called health insurance claim prediction using machine learning seeks to increase the precision and effectiveness of processing insurance claims. Machine learning models can be trained to recognise patterns and forecast the likelihood that a claim will be approved or refused by utilising historical data and prediction algorithms.

This can improve customer satisfaction, reduce fraud, and simplify the claims process for insurance companies. Machine learning can also assist medical personnel in identifying those with a likelihood of getting particular illnesses and offering them preventative therapy, improving patient outcomes.

The application of machine learning in health insurance claim prediction is becoming increasinglycrucial for the industry to remain competitive and deliver high-quality care to patients as healthcare data continues to grow and get more complex.

The purpose of utilising machine learning to forecast health insurance claims is to increase the precision and effectiveness of processing insurance claims in the healthcare sector. Scikit, Numpy, Pandas, and Tensorflow are a some machinelearning software packages that can be used to create this system.

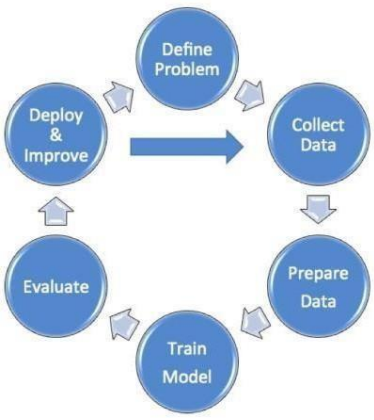


Fig.1.
Workflow

## 3. RELATED WORKS

Medical insurance claim prediction using ml algorithms is an active area of research due to its potential to improve the accuracy and efficiency of claim processing. This is a summary of a few recent research efforts in this area.

A systematic review of "Health Insurance Claim Prediction Using Machine Learning Algorithms" by S. Khan et al. This investigation performed a systematic analysis of 32 research that predicted healthinsurance claims using machine learning algorithms. The most widely utilised algorithms, according to theauthors, are decision trees, logistic regression, and neural networks.

The article "Predicting Health Insurance Claims Using Machine Learning Techniques: A Comparative Analysis" was written by D. P. Shukla et al. In order to anticipate health insurance claims, this study assessed the effectiveness of four machine learning algorithms In terms of accuracy, the authors discovered that random forest fared better than the other algorithms.

In "Health Insurance Claim Prediction Using Machine Learning Algorithms: A Comparative Analysis," S. P. Singh et al. Six machine learning algorithms were evaluated in this study for their ability to predict health insurance claims. In terms of accuracy, the team discovered that random forest and support vector machine fared better than the other algorithms.

The article "Health Insurance Claim Prediction Using Deep Learning Approaches" was written by S. K. Jha et al. In order to forecast health insurance claims, this study applied deep learning techniques. The convolutional neural network functioned more accurately than the long short-term memory, according to the authors.

Hybrid Machine Learning Methods for Health Insurance Claim Prediction, S. Sharma et al. In order to anticipate health insurance claims, this study developed a hybrid machine learning strategy that includes decision trees, k-nearest neighbours, and random forests.

The hybrid technique performed more accurately than the individual algorithms, according to the authors.

## 4. MATERIALS AND METHODOLOGY

The dataset was downloaded in csv format from Kaggle. There are 6 columns of prediction featuresin this dataset's 1338 records. Both categorical and continuous data are present in this dataset. The dataset's characteristics are listed below.

| Features | Representation |
|---|---|
| Age | Age of the patient |
| Sex | Gender of the patient |
| Children | Number of children to the patient |
| B.M.I | Body mass index of the patient |
| Region | Residential area of the patient |
| Smoker | Smoking habits of the patient |
| Charges | Medication cost of the patient |

TABLE.1. Features of Dataset

There are no missing or null values in the insurance dataset that was used for this study. Hence, there is not much preprocessing to be done. However, to turn the categorical data into continuous data, we applied label encoding. Next, we must use different regression models, including linear regression, random forest regression, and decision tree regression, to train our dataset. Several evaluation metrics, including Mean Squared Error (MSE), R2 score, and Root Mean Squared Error (RMSE), are used to compare the training dataset (MSE). Themost accurate model will be selected to forecast the user's health insurance claims.

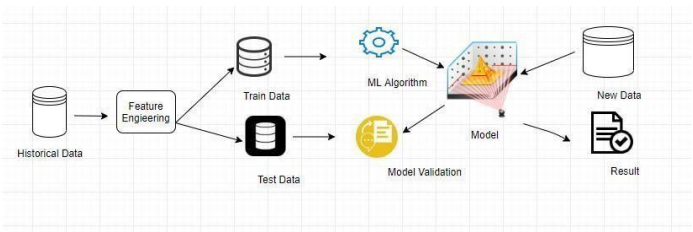


Fig.2. Workflow of insurance cost prediction

To better comprehend the dataset, we have done a graphical depiction. Here are some examples of data visualisation.

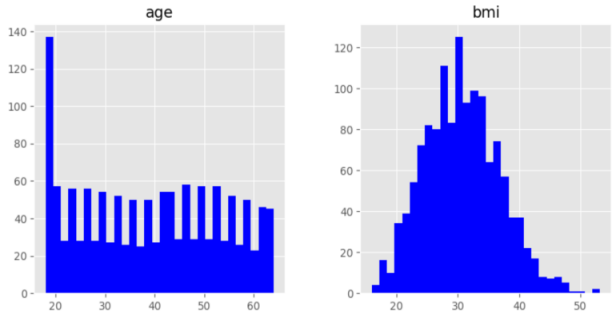


Fig.3. Age and BMI

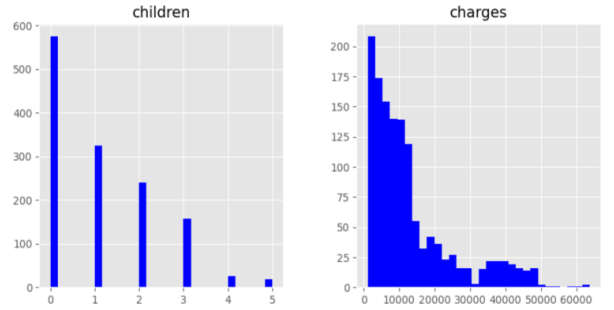


Fig.4. Children and Charges

## 5. REGRESSION METHODS

Regression methods are used in statistics to examine the connection of both a dependent variable and one or even more independent variables. Regression analysis is a powerful tool that can shed light on the underlying connections and patterns that the information reveals whenused in research, forecasts, and decision-making.

To train the dataset for this system, we used three regression techniques: linear regression, random forest regression, and decision tree regression.

### A. *RANDOM FOREST REGRESSION*

The random forest regression approach works by building a number of decision trees, each trained on a different subset of data that is chosen at random. Using the values of the independent variables as their basis, the trees are made todivide the data into more manageable and homogeneous groupings. Thefinal forecast is the average of all individual tree predictions. Each tree makes a prediction based on the mean ormode of the dependent variable inside its terminalnodes.

The ability of random forest regression to handle high-dimensional datasets and datasets with a lot of linked variables is one of its key benefits. In order to choose features and understand models, the algorithm can determine which variables are most crucial for prediction. Also, compared toconventional regression models, random forest regression is less prone to overfitting, which can be problematic when working with complicated data.

Random forest regression does have certain drawbacks, though. When dealing with huge datasets, the approach can be computationally demanding. Moreover, as the technique needs complete data for training, it is not suitable for datasets containing incomplete data.

## 6. RESULTS

The performance of the several regressors were analysed in order to pick the most effective fundamental algorithm and establish the correctness and reliability of the model's deployment. The efficiency of the regressor can be explained using the estimation coefficient and the Root Mean Squared Error (RMSE) (r2 score).

| MODEL | R2 SCORE | RMSE | ACCURACY |
|---|---|---|---|
| Linear Regression | 0.7694415 | 6144.199 | 76 % |
| Random Forest Regression | 0.8525963 | 1876.602 | 85 % |
| Decision Tree Regression | 0.7293854 | 6299.100 | 72 % |

TABLE.2. Comparison of models

Comparing the multiple regression models used in Table 2 for predicting the insurance claim revealed that the larger the r-squared value of the models. Hence, the R-squared number gives the most accurate representation of how variable the dependentvariables is in relation to the surrounding mean.

With an R2 value of 0.8525963, the Random Forest Regression model was determined to be the best model.

## CONCLUSION AND FUTURE WORK

The expense of the health policy has been estimated as precisely as possible using the predictive models described here. To create better medical facilities, this is allegedly highly beneficial for the healthcare organisation. The study made use of Decision Tree Regression, Random Forest Regression, and Linear Regression methodologies. When all the results were compared, Random Forest Regressor had the greatest R2 score.

Future research, however, can focus on exploring a number of additional deep learning applications. The technique that can converge the changes and do multiple framework can be improved using improved approaches from machine learning and other fields. Although most forecast errors involve a significant financial claim transaction, updating some of them can generally benefit insurance firms.

## REFERENCES


[1] "Prediction of Insurance Claim Severity Loss Using Regression Models," R. M. Ogunnaike and D. Si, 2017, pp. 233-247.

[2] "Modeling frequency and severity of claims with the zero-inflated generalised cluster weighted models," Insur. Math. Econ., vol. 94, pp. 79–93, September 2020; doi: 10.1016/j.insmatheco.2020.06.004..

[3] "A complete overview and analysis of generative models in machine learning," Computer Science Review, vol. 38, no. 100285, 2020, doi: 10.1016/j.cosrev.2020.100285.

[4] "Nuclei segmentation in cell pictures using fully convolutional neural networks," Int. J. Emerg. Technol., vol. 11, no. 3, pp. 731–737, 2020. S. S. Rautaray, S. Dey, M. Pandey, and M. K. Gourisaria.

[5] Evaluation of Technology Adoption Models and Theories to Assess Readiness and Appropriate Usage of Technology in a Corporate Organization by T. Dube, R. Van Eck, and T. Zuva 10.36548/jitdw.2020.4.003

[6] T. J. Layton, "Imperfect risk adjustment, risk preferences, and sorting in competitive healthinsurance markets," J. Inf. Technol. Digit. World, vol. 02, no. 4, pp. 207–212, 2020. 56, 259–280, J. Health Econ., 2017, doi: 10.1016/j.jhealeco.2017.04.004.

[7] Jayasree, M., and M. G. Chandrasekhar (2020). Machine Learning Methods for Predicting Health Insurance Claims. The International Conference on Communication and Signal Processing (ICCSP) will take place in 2020. (pp. 0426-0431). IEEE.

[8] P. Awasthi, P. Sharma, and S. (2020). Machine Learning for Predictive Analysis of Health Insurance Claims. The eleventh ICCCNT (International Conference on Computation, Communication, and Networking Technologies)will take place in 2020. (pp. 1-6). IEEE

[9] Albarrak, A. M., Elshaikh, E. A., and Basheer, M. A. (2021). A thorough review of the prediction of health insurance claims using machine learning techniques. 45(1), 1-14 in Journal of Medical Systems.

# BG6

ORIGINALITY REPORT

| 4% | 4% | 4% | 0% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | www.ijcaonline.org<br>Internet Source | 2% |
| 2 | Hui Li Tan, Zhengguo Li, Yih Han Tan, S. Rahardja, Chuohuo Yeo. "A Perceptually Relevant MSE-Based Image Quality Metric", IEEE Transactions on Image Processing, 2013<br>Publication | 1% |
| 3 | www.simplilearn.com<br>Internet Source | 1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On | | |

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

**PAPER ID**
NECICAIEA2K23025

International Conference on
# Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

# Certificate of Presentation

This is to Certify that Bolla Jhansi Vazram , Narasaraopeta engineering college has presented the paper title Health insurance claims prediction using machine learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

**Convenor**
**Dr.S.V.N.Srinivasu**

**Chief-Convenor**
**Dr.S.N.Tirumala Rao**

**Principal, Patron**
**Dr. M. Sreenivasa Kumar**

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

**PAPER ID**
NECICAIEA2K23025

International Conference on
# Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

# Certificate of Presentation

This is to Certify that Bandi Kalyani , Narasaraopeta engineering college has presented the paper title Health insurance claims prediction using machine learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

**Convenor**
**Dr.S.V.N.Srinivasu**

**Chief-Convenor**
**Dr.S.N.Tirumala Rao**

**Principal, Patron**
**Dr. M. Sreenivasa Kumar**

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

**PAPER ID**
NECICAIEA2K23025

International Conference on
# Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

# Certificate of Presentation

This is to Certify that Kolisetty Lalitha Annapurna , Narasaraopeta engineering college has presented the paper title Health insurance claims prediction using machine learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

**Convenor**
**Dr.S.V.N.Srinivasu**

**Chief-Convenor**
**Dr.S.N.Tirumala Rao**

**Principal, Patron**
**Dr. M. Sreenivasa Kumar**

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in

**PAPER ID**
NECICAIEA2K23025

International Conference on
# Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI

# Certificate of Presentation

This is to Certify that Ambati Bhavani , Narasaraopeta engineering college has presented the paper title Health insurance claims prediction using machine learning in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.

**Convenor**
**Dr.S.V.N.Srinivasu**

**Chief-Convenor**
**Dr.S.N.Tirumala Rao**

**Principal, Patron**
**Dr. M. Sreenivasa Kumar**