PROJECT TITLE :- MEDICAL INSURANCE COST PREDICTION

1. INTRODUCTION

1.1 OVERVIEW

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 numerous products to shield individuals and organisations from these risks by using financial capital to shield them. Therefore Insurance is one of the policies that either decreases or removes loss costs incurred by various risks.

Health insurance costs have risen dramatically over the past decade in response to the rising cost of health care services and are determined by a multitude of factors. Let's look at the cost of healthcare for a sample of the population given age, sex, bmi, number of children, smoking habits, and region. Seek insight from the dataset with Exploratory Data Analysis

• Performed Data Processing, Data Engineering and Feature Transformation to prepare data before modelling• Built a model to predict Insurance Cost based on the features• Evaluated the model using various Performance Metrics like RMSE, R2, Testing Accuracy, Training Accuracy and MAE

1.2 PURPOSE

The purpose of this project is to determine the contributing factors and predict health insurance cost by performing exploratory data analysis and predictive modelling on the Health Insurance dataset. This project makes use of Numpy, Pandas, Sci-kit learn, and Data Visualization libraries.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM

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 victimisation 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. 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 center 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 centimeter methods that are extremely applicable for the issues of many missing values, such as XGboost.

G. Kowshalya, M. Nandhini. in 2018 classifiers are developed during this study to predict and estimate dishonorable claims and a proportion of premiums for the varied customers based mostly upon their personal and monetary data. For classification, the algorithms Random Forest, J48, and Naïve Bayes are chosen. The findings show that Random Forest exceeds the remaining techniques betting on the artificial dataset. This paper thus doesn't cowl claim forecasts, however rather focuses on false claims . The on top of previous works failed to contemplate each foreseen the value or claim severity, they solely create a classification for the issues of claims (whether or not a claim was filed for that policyholder) during this study we tend to specialize in advanced applied math ways and machine learning algorithms and deep neural network for predict the value of health insurance.

2.2 PROPOSED SOLUTION

A.Multiple Linear Regression.

Multiple regression may be an applied mathematics technique which will be wont to analyze the link between one variable and a number of other freelance variables. For example, with the information set utilized in this study, we might need to grasp independent variables (8 independent variables), (linearly) involving the dependent variable (charges). is} mentioned because of the multiple simple regression (MLR) model. associate degree MLR model with t\ independent options and Y results can be calculated as within the following equation

In the higher than equation, u is that the residual regression whereas ? is the weight of every independent variable or parameter assigned.

B. Decision trees

DTs are straightforward, terribly popular, fast-training, and straightforward to scan models with comparative or different strategies of learning from the data. they're fairly competent however prone to overfitting in their predictions. they'll be strong by their performance .

C. Gradient Boosting Regression

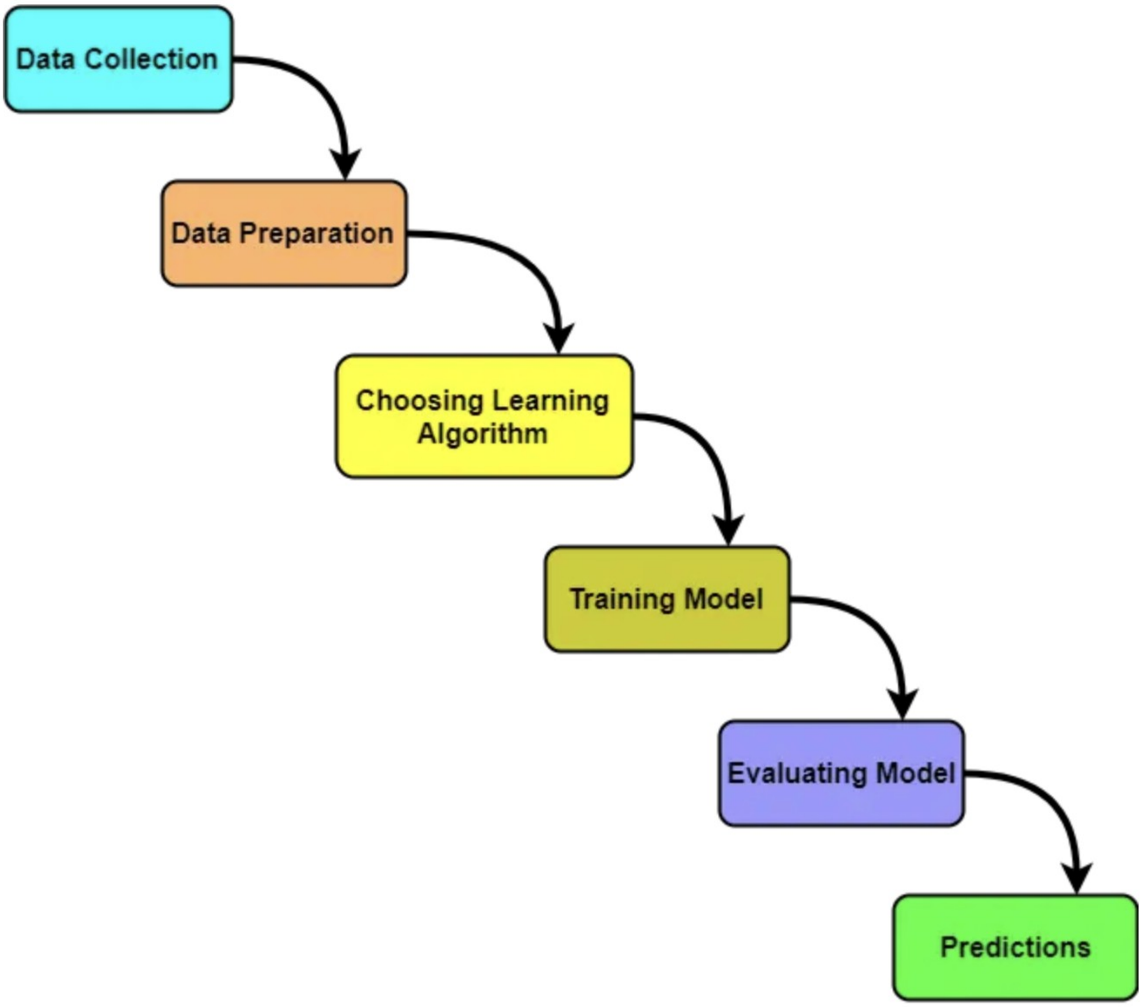
Gradient boosting algorithmic program is one among the foremost powerful algorithms within the field of machine learning. As we all know that the errors in machine learning algorithms are generally classified into 2 classes i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it's accustomed to minimize bias error of the model .

Gradient boosting algorithms are often used for predicting not solely continuous target variables as a regression however additionally categorical target variables (as a Classifier). Once it is used as a regressor, the price operates as Mean square. Error (MSE) and when it is used as a classifier then the price operates as Log loss.

3. THEORETICAL ANALYSIS

3.1 BLOCK DIAGRAM

The diagram represents the implementation of the project that has done step by step which has data preprocessing, cleaning, training and testing of the data as well.



Data Collection: Gather relevant data about individuals' attributes (age, gender, BMI, smoking habits, etc.) and their corresponding medical insurance costs.

Data Preprocessing: Clean the collected data by handling missing values, outliers, and transforming categorical variables into numerical representations.

Feature Selection: Identify the most significant features that impact medical insurance costs using techniques like correlation analysis or domain knowledge.

Model Development: Choose an appropriate machine learning algorithm, such as multiple linear regression, decision trees, random forests, or neural networks, for building the medical insurance cost prediction model.

Model Training: Train the selected model using the preprocessed data, allowing it to learn the underlying patterns and relationships between the features and the target variable.

Model Evaluation: Assess the performance of the trained model using evaluation metrics like mean squared error (MSE), mean absolute error (MAE), or R-squared to measure its accuracy and reliability.

Model Optimization: Fine-tune the model by adjusting hyper parameters, employing regularisation techniques, or exploring ensemble methods to enhance its performance.

Prediction: Deploy the optimized model to make predictions on new, unseen data, providing the relevant attributes as input to estimate the medical insurance costs for individuals.

This block diagram provides a high-level overview of the project, highlighting the major steps involved in developing a medical insurance cost prediction system.

3.2 HARDWARE/SOFT WARE COMPONENTS

Hardware components

System: Pentium IV 2.4 GHz ← Hard Disk:40GB ← Floppy Drive: 1.44Mb ← Monitor: 15VGA Color ← Mouse: Logitech ← Ram:512Mb

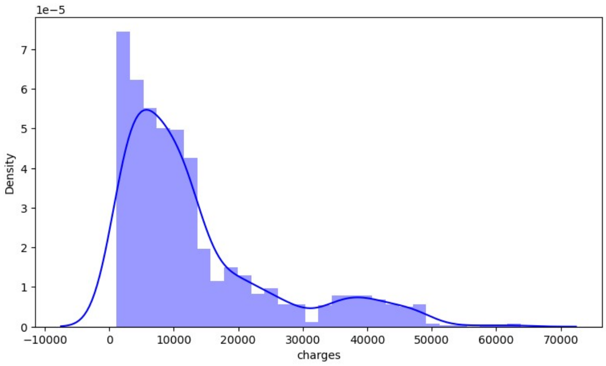
Software components

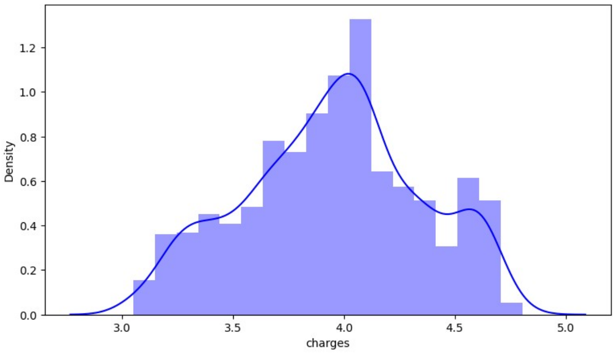
← Operating system: Windows XP/7 ← Coding Language: python ← IDE: Anaconda Navigator

4.EXPERIMENTAL INVESTIGATION

Exploratory Data Analysis :

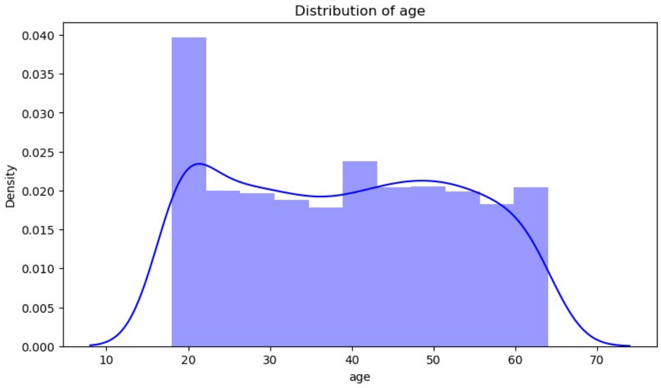
Here we can see some analysis made while working on the project.

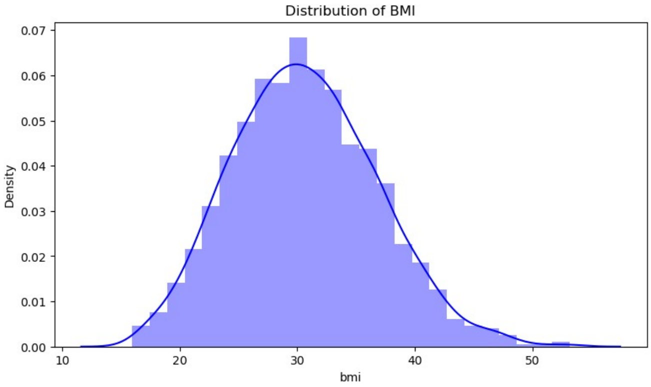




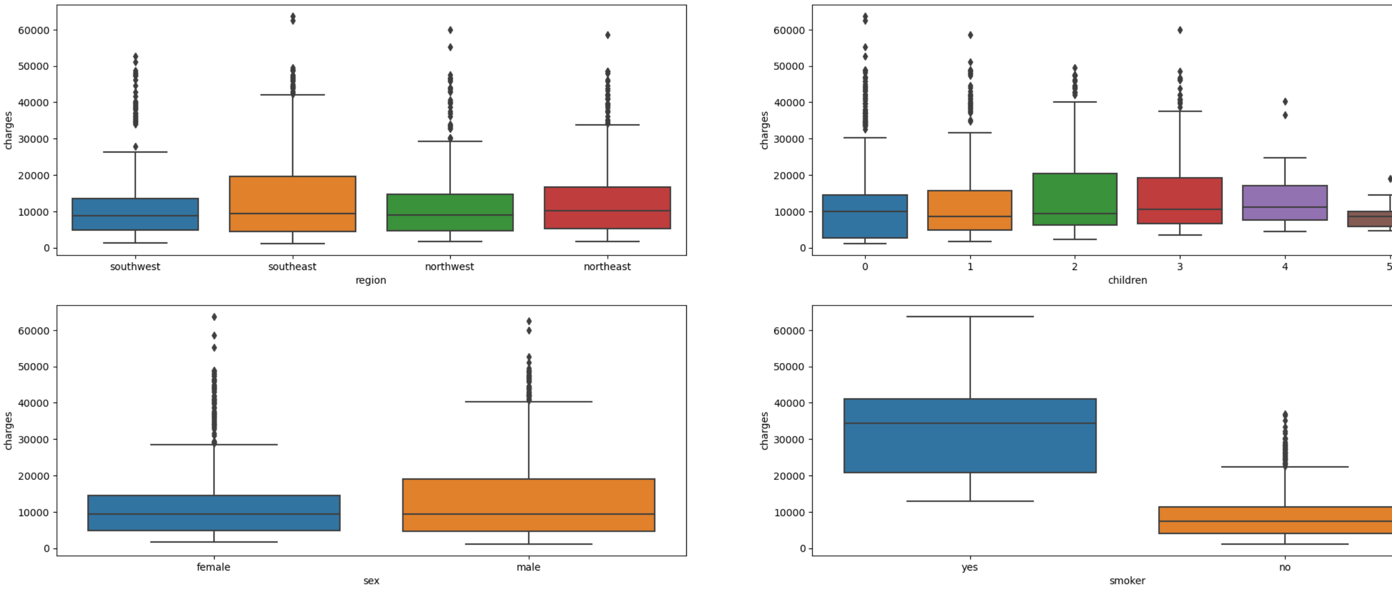
The above figure show how the charges are distributed according to the charge range and density of how much people are range of certain charge.

Visualizing distribution of numerical variables vs medical charges :





Visualizing categorical variables individually vs medical charges:

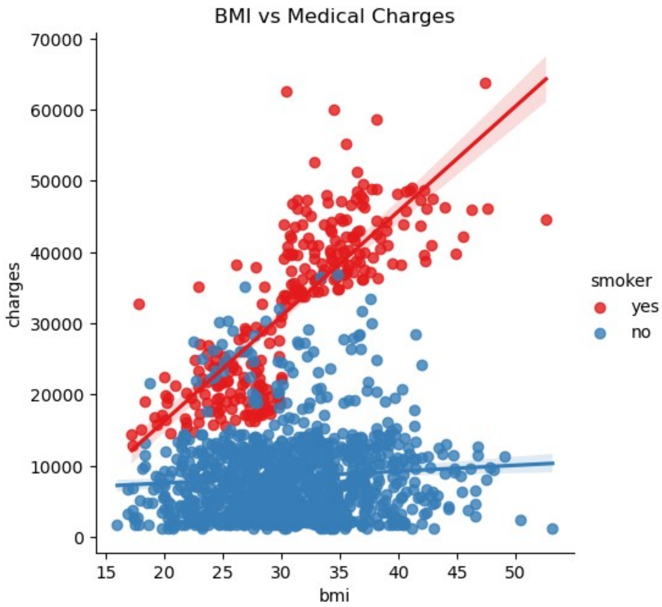


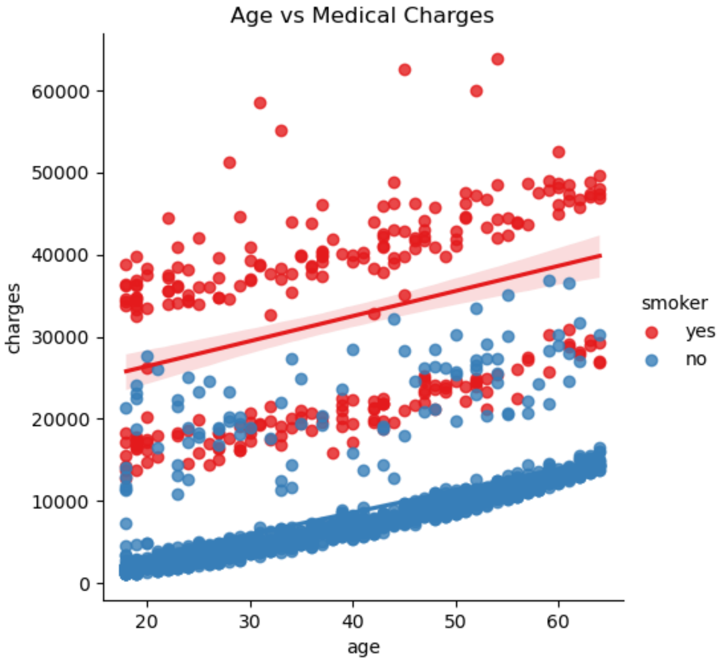
Charges by age, bmi, and children based on smoking behaviour :

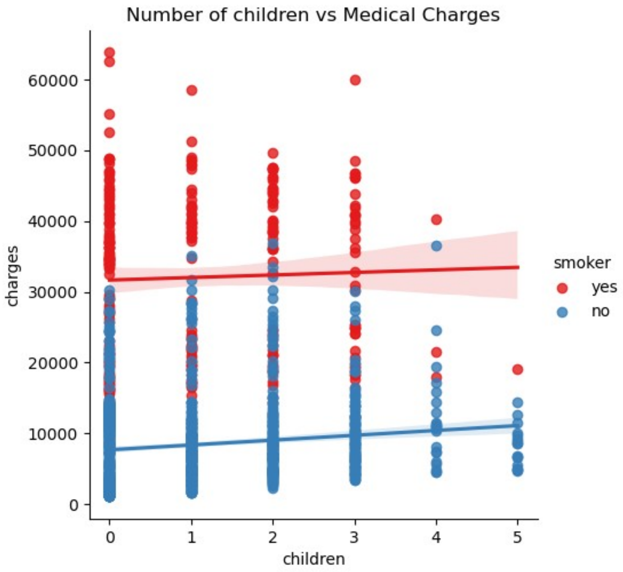
Based on the analysis,

We know that smoking has a great impact on medical costs.

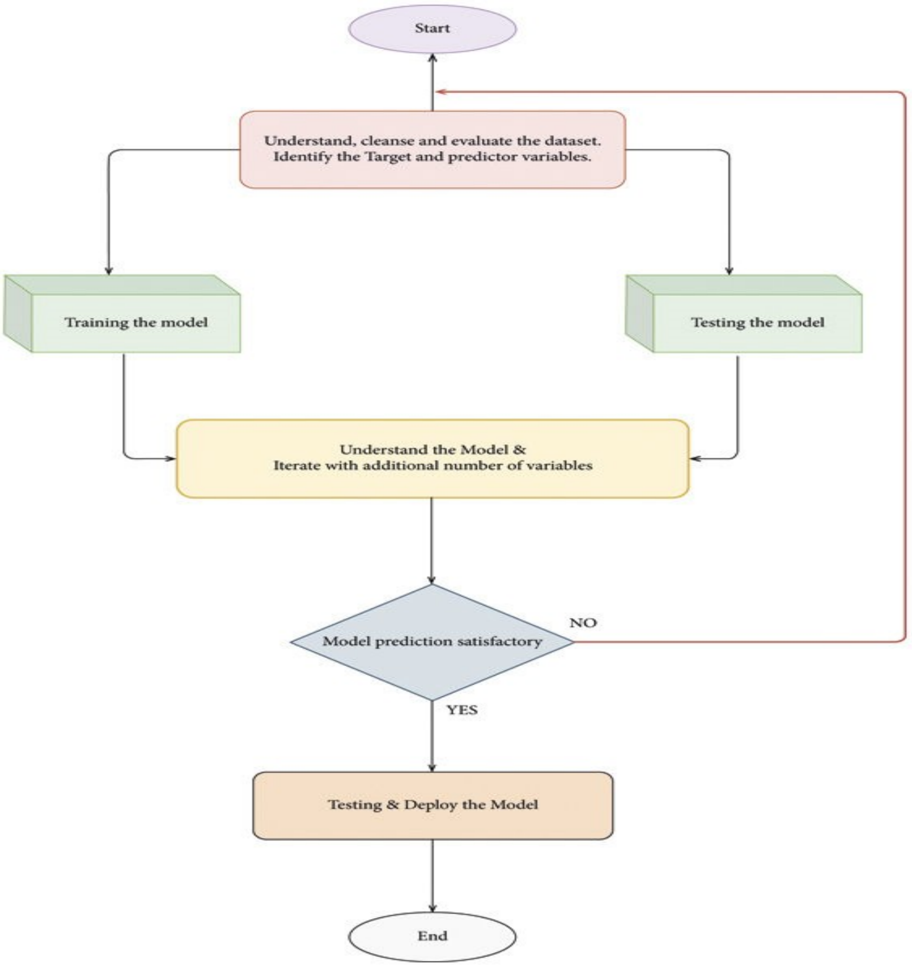
Smoking in combination with an increase in other factors, such as age, bmi, and children, further emphasizes and increases the medical cost due to smoking.


BMI vs Medical Charges
charges
70000
60000
50000
40000
30000
20000
10000
0
15
20
25
30
35
40
45
50
bmi
smoker
yes
no


Age vs Medical Charges
charges
60000
50000
40000
30000
20000
10000
0
20
30
40
50
60
age
smoker
yes
no


Number of children vs Medical Charges
60000
50000
40000
30000
20000
10000
0
charges
0
1
2
3
4
5
children
smoker
yes
no

5.FLOW CHART :

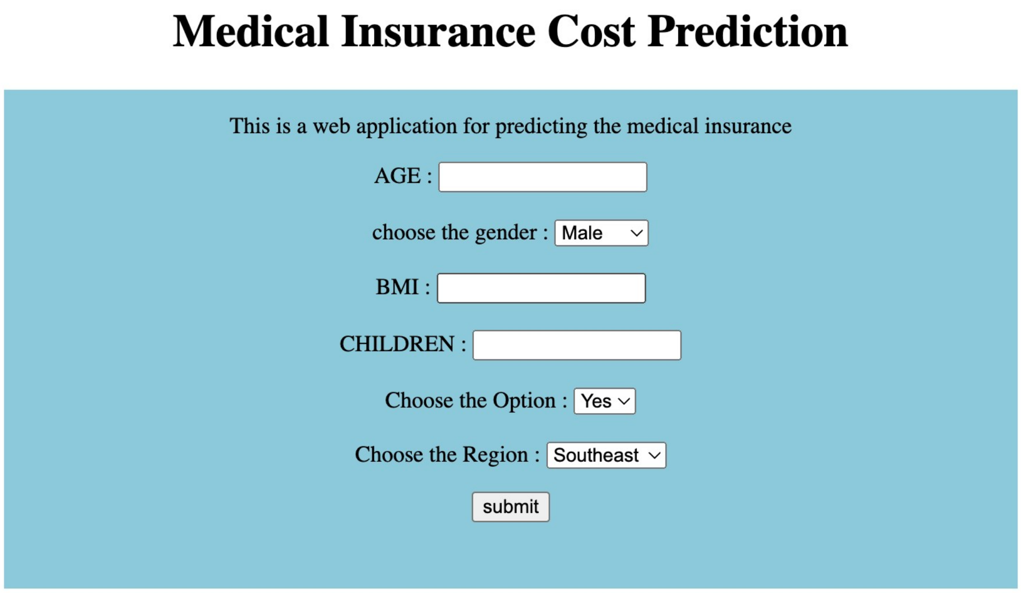

Start
Understand, cleanse and evaluate the dataset.
Identify the Target and predictor variables.
Training the model
Testing the model
Understand the Model &
Iterate with additional number of variables
Model prediction satisfactory
NO
YES
Testing & Deploy the Model
End

6. RESULT :

There are many algorithms can be used for predicting the any model we will try all algorithms on our model and deploy the algorithm which we are getting predicted value as nearer to the actual value we want .

So using linear regression algorithm we had deployed our model in flask to get the predicted value for the given attributes in html webpage , so using the predicted value we can compare the actual value present in dataset to find how much accurate the model is

.



Medical Insurance Cost Prediction

This is a web application for predicting the medical insurance

AGE :

choose the gender : Male

BMI :

CHILDREN :

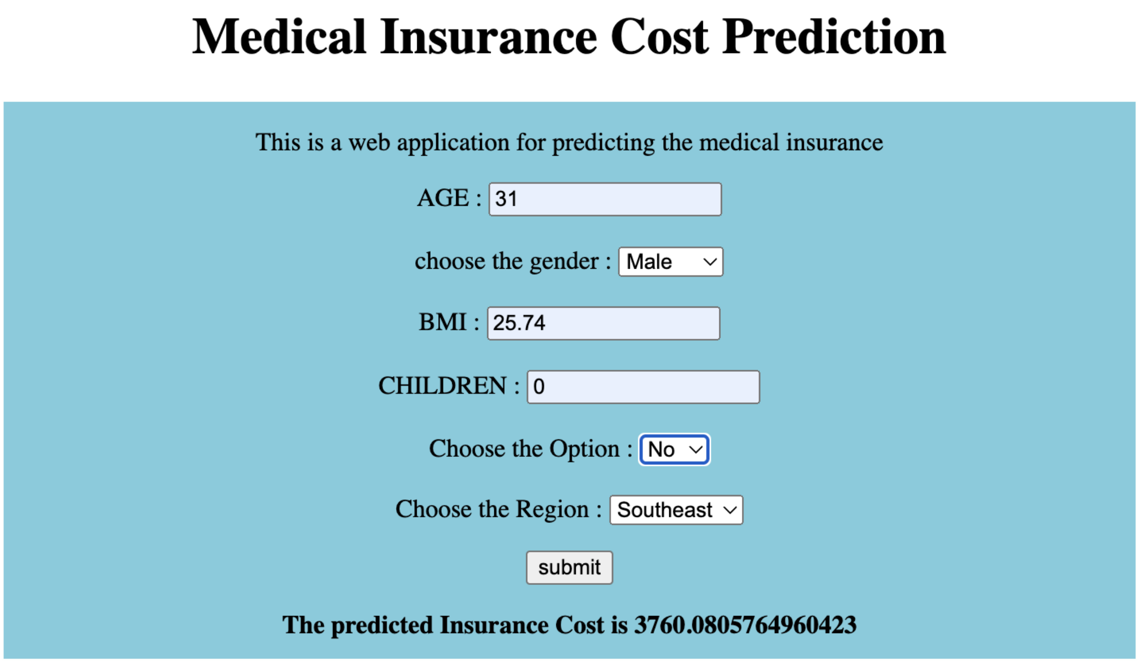
Choose the Option : Yes

Choose the Region : Southeast

This is the view of the html page we had created in flask now let us try to give input and find the predicted value.

Here we will give input as 31,1,25,74,0,1,0 :

For the objects we can directly the value according to dataset information but we had changed the categorical features (ex: smoker {yes : 0, no : 1} , gender{male:0, female:1} and region {southeast : southwest:1, northeast : 2 , northwest: 3}) so according to this values we have to give the input .



Medical Insurance Cost Prediction

This is a web application for predicting the medical insurance

AGE : 31

choose the gender : Male

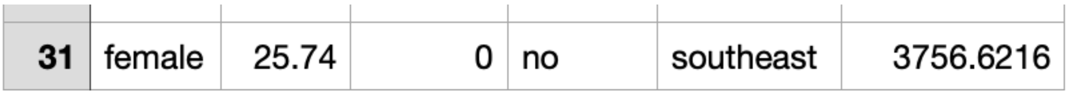
BMI : 25.74

CHILDREN : 0

Choose the Option : No

Choose the Region : Southeast

The predicted Insurance Cost is 3760.0805764960423



31	female	25.74	0	no	southeast	3756.6216

The actual value we need for the given input is 3756.6216 we are getting the value 3760.080 means we are getting predicted value as very nearer to the accurate value among all the algorithms we had used we are getting more accurate value while using linear regression , you can check for any input in data set by running the python file which we deployed using flask

7.ADVANTAGES AND DISADVANTAGES :

Linear regression offers several advantages for predicting medical insurance costs:

Simplicity: Linear regression is a straightforward and easy-to-understand method for predicting costs. It assumes a linear relationship between the predictor variables (e.g., age, BMI, smoking status) and the target variable (insurance cost), making it a simple and intuitive approach.

Interpretability: With linear regression, the coefficients of the predictor variables provide insight into their individual impact on the insurance cost. Positive coefficients indicate that an increase in the predictor variable leads to an increase in the cost, while negative coefficients suggest a decrease in the cost.

Speed and Efficiency: Linear regression models are computationally efficient and can handle large datasets relatively quickly. This makes it suitable for predicting medical insurance costs, where there may be a considerable number of data points and predictor variables.

Baseline Model: Linear regression serves as a useful baseline model for comparison against more complex algorithms. It provides a simple benchmark to evaluate the performance of other predictive models.

Feature Selection: Linear regression helps identify the most important predictor variables for predicting insurance costs. By analyzing the magnitude and significance of the coefficients, you can determine which variables have the most significant impact on the outcome.

Linear Assumption: Although the linear assumption may not hold in all cases, it can still provide a reasonable approximation for many real-world scenarios. If the relationship between predictors and the target variable is approximately linear, linear regression can yield accurate predictions.

No Distribution Assumption: Linear regression does not require any assumptions about the distribution of variables, except for the target variable. This flexibility allows for broader application across various datasets without specific distribution requirements.

Model Transparency: Unlike complex models such as neural networks or ensemble methods, linear regression models are transparent and offer clear insight into the prediction process. This transparency is beneficial for understanding the factors influencing insurance costs and gaining stakeholders' trust in the model's predictions.

While linear regression has its advantages, it's important to note that it may not capture complex nonlinear relationships between predictors and insurance costs. In such cases, more advanced techniques like polynomial regression or machine learning algorithms could be explored.

Disadvantages :

While linear regression has its advantages, there are also some limitations and disadvantages when using it for predicting medical insurance costs:

Linearity Assumption: Linear regression assumes a linear relationship between the predictor variables and the target variable. However, in reality, the relationship between insurance costs and predictors may not be strictly linear. Complex nonlinear relationships may exist, and linear regression may fail to capture them accurately.

Limited Flexibility: Linear regression is limited in its ability to model complex interactions and nonlinear effects. If there are intricate interactions or nonlinearities present in the data, linear regression may not provide an accurate prediction.

Outliers and Influential Points: Linear regression is sensitive to outliers and influential points, which can distort the model's performance and affect the predictions significantly. If the data contains extreme values or influential observations, linear regression may not produce robust results.

Multicollinearity: Linear regression assumes that predictor variables are not highly correlated with each other. When multicollinearity exists, meaning high correlation between predictors, it can cause issues such as unstable coefficient estimates and difficulty in interpreting the individual effects of predictors.

Limited Feature Representation: Linear regression assumes a linear relationship between predictors and the target variable. It may struggle to capture complex interactions, non-additive effects, and higher-order relationships. In medical insurance cost prediction, there might be complex interactions between multiple predictors that linear regression cannot adequately model.

8.APPLICATIONS :

Some specific applications where the solution can be applied are :

Individual Insurance Premiums: Linear regression can be employed to estimate the insurance costs for individual policyholders based on their personal characteristics such as age, gender, BMI, smoking status, pre-existing conditions, and other relevant factors. By fitting a linear regression model to historical data, insurers can predict the premiums for new policyholders.

Claims Analysis: Linear regression can help analyze the relationship between medical claims and associated costs. By examining variables such as diagnosis, treatment procedures, hospitalization duration, and other factors, linear regression can identify patterns and estimate the expected costs of different types of claims. This information can aid insurers in setting appropriate coverage levels and pricing policies.

Risk Assessment: Linear regression can be utilized to assess the risk profile of policyholders based on their characteristics. By analyzing data on factors such as age, medical history, lifestyle, and occupation, insurers can build regression models to predict the potential cost of insuring an individual or a group. This enables insurers to classify policyholders into risk categories and determine appropriate premiums.

Cost Comparisons: Linear regression can be used to compare the costs of different medical treatments or interventions. By examining patient outcomes, treatment options, and associated costs, linear regression can estimate the cost-effectiveness of various healthcare interventions. This information is valuable for insurers, healthcare providers, and policymakers in making informed decisions regarding coverage and reimbursement.

Premium Adjustments: Linear regression models can be employed to adjust insurance premiums based on changes in relevant factors. For example, if there is an increase in medical inflation or changes in the population's health status, linear regression can help insurers evaluate the impact on insurance costs and make appropriate adjustments to premiums.

9.CONCLUSION

In conclusion, linear regression is a useful method for predicting medical insurance costs in various applications. It offers simplicity, interpretability, and efficiency, making it a practical choice for estimating premiums, analyzing claims, assessing risk, and making underwriting decisions. Linear regression provides a baseline model for comparison and helps identify important predictor variables. However, it has limitations, such as the assumption of linearity, sensitivity to outliers, and difficulty capturing complex interactions or non-linear relationships. Careful consideration of these limitations is necessary when applying linear regression to medical insurance cost prediction. Additionally, it's important to note that more advanced techniques may be required in cases where linear regression falls short. Overall, linear regression serves as a valuable tool in the field of medical insurance cost prediction, offering insights into factors influencing costs and supporting decision-making processes for insurers, healthcare providers, and policymakers.

10.FURTUE SCOPE

There are several potential enhancements that can be made in the future to improve the model for medical insurance cost prediction:

1. Nonlinear Regression: As mentioned earlier, linear regression assumes a linear relationship between predictors and the target variable. Exploring nonlinear regression models, such as polynomial regression or generalized additive models, could capture more complex relationships and improve the accuracy of predictions.

2. Feature Engineering: Feature engineering involves creating new features or transforming existing ones to enhance the predictive power of the model. By incorporating domain knowledge and considering interactions, polynomial terms, logarithmic transformations, or interaction terms, the model can better capture the complexities of medical insurance costs.

3. Advanced Machine Learning Algorithms: While linear regression is a valuable tool, more advanced machine learning algorithms, such as decision trees, random forests, gradient boosting, or neural networks, can potentially capture complex patterns and interactions in the data. These algorithms have the ability to handle nonlinear relationships and automatically select relevant features, leading to improved predictions.

4. Model Ensemble: Combining the predictions of multiple models, known as model ensemble, can often improve the overall performance. Ensembling techniques, such as stacking, bagging, or boosting, can be

applied to linear regression or a combination of different regression models to leverage their strengths and mitigate their weaknesses.

It's worth noting that the choice of enhancements depends on the specific characteristics of the data, the complexity of relationships, and the goals of the insurance company or research study. Evaluating different techniques and selecting the most appropriate enhancements should be done based on rigorous experimentation and thorough evaluation of the model's performance.

11.BIBILOGRAPHY :

https://www.emerald.com/insight/content/doi/10.1108/XJM-07-2020-0021/full/html

https://towardsdatascience.com/how-to-easily-deploy-machine-learning-models-using-flask-b95af8fe34d4

https://www.kaggle.com/code/nishxnt/medical-insurance-cost-prediction-python

https://www.pinterest.com/pin/789748484662037894/

APPENDIX :

Source Code :

App.py file :

```
from flask import Flask , render_template , request

app = Flask(__name__) # interface between my server and my application

import pickle

model = pickle.load(open('/Users/sanjaytulabandula/Desktop/DataScienceProject/model.pkl' , 'rb'))


@app.route('/')# binds to an url

def helloworld():
    return render_template("index.html")


@app.route('/login' , methods = ['POST'])# binds to an url
def login():
    p = request.form["ag"]
    s = request.form["s"]


    b = request.form["bm"]
    c = request.form["ch"]
    sm = request.form["sm"]


    rg = request.form["rg"]




    t=[[int(p),int(s),float(b),int(c),int(sm),int(rg)]]
    output= model.predict(t)
    print(output)


    return render_template("index.html" , Y = "The predicted Insurance Cost is " + str(output[0]))
if __name__ == '__main__' :
    app.run(debug= True)
```

index.html :

```
<html>
  <head>
    <center><h1>Medical Insurance Cost Prediction</h1></center>
```

```
   <style media="screen">
     span{
       color:red;

  }
    .div1{
     background-color: rgba(110, 201, 219, 0.939);
      border:1px solid rgb(253, 255, 253);
      width:50%;
      height:350px;
      background-position: center;
    }
    main{
      background-color: aquamarine;
    }

  </style>

</head>

  <center>

   <div class="div1">
     <p><center>This is a web application for predicting the medical insurance</center> </p>
     <center>
         <form action = "/login" method = "post">
```

```html
<center><label for="age">AGE : </label>
    <input type="number" name="ag" ></center>  <br>

<label for = "sex" >  choose the gender :  </label>
<select name = "s">
<option Value = "0">Male</option>
<option Value = "1">Female</option>
</select>   <br>

<br>

<center><label for="bmi">BMI : </label>
    <input type="text" name="bm" ></center>  <br>



    <center><label for="children">CHILDREN : </label>
        <input type="text" name="ch" ></center>  <br>



<label for = "smoker" >  Choose the Option :  </label>
<select name = "sm">
<option Value = "0">Yes</option>
<option Value = "1">No</option>
</select>

<br>
<br>
```

```
            <label for = "region" >  Choose the Region : </label>
            <select name = "rg">

            <option Value = "0">Southeast</option>
            <option Value = "1">Southwest</option>
            <option Value = "2">Northeast</option>
            <option Value = "3">Northwest</option>

            </select>



            <p><input type = "submit" value = "submit" /> </p>

        </form>
        <b>{{Y}}</b>
    </center>
 </div>
  </center>

</body>
</html>
```

The remaining Jupyter files can be accessed with the below GitHub link of my profile , you can find all the files related to model in the repository created in the profile

Github link : https://github.com/sanjay6621/Medical-Insurance-Cost-Prediction.git

Thank you