VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

A Mini Project Report on

"MEDICAL EXPENDITURE PREDICTION SYSTEM"

Submitted in partial fulfillment of the requirements as a part of the

AI/ML INTERNSHIP

(NASTECH)

For the award of degree of

Bachelor of Engineering
in
Information Science and Engineering

Submitted by

SAMARTH R AITHAL — **1RN18IS091**

SANDESH A RAM — **1RN18IS093**

Internship Project Coordinators

Dr. R Rajkumar	**Mr. Santhosh Kumar**
Associate Professor	**Assistant Professor**
Dept. of ISE, RNSIT	**Dept. of ISE, RNSIT**

Department of Information Science and Engineering

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,
Bengaluru – 560 098

2020 -2021

RNS Institute of Technology

Channasandra, Dr. Vishnuvardhan Road, RR Nagar Post,

Bengaluru – 560 098

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the mini project report entitled ***MEDICAL EXPENDITURE PREDICTION SYSTEM*** has been successfully completed by **SAMARTH R AITHAL** bearing USN **1RN18IS091** and **SANDESH A RAM** bearing USN **1RN18IS093** , presently VII semester students of **RNS Institute of Technology** in partial fulfillment of the requirements as a part of the ***AI/ML Internship (NASTECH)*** for the award of the degree of ***Bachelor of Engineering in Information Science and Engineering*** under **Visvesvaraya Technological University, Belagavi** during academic year **2021 – 2022**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report and deposited in the departmental library. The mini project report has been approved as it satisfies the academic requirements as a part of Internship.

Dr. R Rajkumar	**Mr. Santhosh Kumar**	**Dr. Suresh L**
Coordinator	Guide	Professor and HoD
Associate Professor	Assistant Professor	

External Viva

Name of the Examiners	**Signature with date**
1.	
2.	

ABSTRACT

The medical sector is one of the most important industries with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve efficiency of working in this industry.

Medical Expenditure prediction project focuses on providing an estimate on the health insurance for a particular person. By analyzing certain conditions such as a person's age, bmi, sex, number of children, whether the person smokes or not, speculated prices will be estimated for the health insurance of that person. The motive of this project is to help the customers to estimate the health insurance price for their family. Some of the related factors that impact the cost were also taken into considerations such as physical conditions, concept and location etc.

Medical expenditure price prediction on a data set has been done by using gradient booster regression technique. Moreover, this project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The aim of our project is to build a predictive model for estimating the cost of insurance for a family so that the family can be prepared in advance to get insured.

ACKNOWLEDGMENT

The fulfillment and rapture that go with the fruitful finishing of any assignment would be inadequate without the specifying the people who made it conceivable, whose steady direction and support delegated the endeavors with success.

We would like to profoundly thank **Management** of **RNS Institute of Technology** for providing such a healthy environment to carry out this AI/ML Internship Project.

We would like to express our thanks to our Principal **Dr. M K Venkatesha** for his support and inspired us towards the attainment of knowledge.

We wish to place on record our words of gratitude to **Dr. Suresh L,** Professor and Head of the Department, Information Science and Engineering, for being the enzyme and master mind behind our Internship Project.

We would like to express our profound and cordial gratitude to my Internship Project Coordinators**, Dr. R Rajkumar,** Associate Professor, Department of Information Science and Engineering for their valuable guidance, constructive comments, continuous encouragement throughout the Internship Project and guidance in preparing report.

I would like to express my profound and cordial gratitude to my Faculty in charge M**r.Santhosh Kumar**, Assistant Professor, Department of Information Science and Engineering for his/her valuable guidance in preparing Project report.

We would like to thank all other teaching and non-teaching staff of Information Science & Engineering who have directly or indirectly helped us to carry out the Internship Project.

Also, we would like to acknowledge and thank our parents who are source of inspiration and instrumental in carrying out this Internship Project Work.

SAMARTH R AITHAL
USN:1RN18IS091

SANDESH A RAM
USN:1RN18IS093

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 ORGANIZATION/INDUSTRY

1.1.1 COMPANY PROFILE

NASTECH is formed with the purpose of bridging the gap between Academia and Industry Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

1.1.2 DOMAIN/TECHNOLOGY

The domain chosen for our project is AI/ML. Machine learning, the fundamental driver of AI, is possible through algorithms that can learn themselves from data and identify patterns to make predictions and achieve your predefined goals, rather than blindly following detailed programmed instructions, like in traditional computer programming. This technology allows the machine to perceive, learn, reason and communicate through observation of data, like a child that grows up and acquires knowledge from examples. Machines also have the advantage of not being limited by our inherent biological limitations. With machine learning, manufacturing companies have increased production capacity up to 20%, while lowering material consumption rates by 4%.

Nowadays, the revolutionary AI technology evolved from rule-based expert systems to machine learning and more advanced subcomponents such as deep learning (learning representations instead of tasks), artificial neural networks (inspired by animal brains) and reinforcement learning (virtual agents rewarded if they made good decisions).

The AI can master the complexity of the intertwining industrial processes to enhance the whole flow of production instead of isolated processes. This enormous cognitive capacity gives the AI the ability to consider the spatial organization of plants and the timing constraints of live production. Another key advantage is the capability of AI algorithms to think probabilistically, with all the subtlety this allows in edge cases, instead of traditional rule-based methods that require rigid theories and a full comprehension of problems.

1.1.3 Department

R.N.Shetty Institute of Technology (RNSIT) established in the year 2001, is the brain-child of the Group Chairman, Dr. R. N. Shetty. The Murudeshwar Group of Companies headed by Sri. R. N. Shetty is a leading player in many industries viz construction, manufacturing, hotel, automobile, power & IT services and education. The group has contributed significantly to the field of education. A number of educational institutions are run by the R. N. Shetty Trust, RNSIT being one amongst them. With a continuous desire to provide quality education to the society, the group has established RNSIT, an institution to nourish and produce the best of engineering talents in the country. RNSIT is one of the best and top accredited engineering colleges in Bengaluru.

1.2 PROBLEM STATEMENT

1.2.1 Existing System and their Limitations

A manual method is currently used in the market to predict the insurance price. The problem with this is that it is very time consuming and tedious process for the insurance company to manually provide this for every customer. To overcome this, insurance companies tend to hire an agent which again increases the cost of the process.

Moreover, there is a chance that the agent might be prone to manual error or bribery.

1.2.2 Proposed Solution

To eliminate the drawback of manual method, Machine learning algorithms can be used by insurance companies to provide a fast, easy and customer friendly approach for the problem. Also, the new system will be cost and time efficient. This will have simple operations.

1.2.3 Program formulation

The proposed system works on Gradient Boosting Regression Algorithm. This algorithm takes into account all the different conditions on which the insurance can be provided and gives a highly accurate estimate.

Chapter 2

REQUIREMENT ANALYSIS, TOOLS &TECHNOLOGIES

2.1 Hardware and Software Requirements

2.1.1 Hardware Requirements:

- Processor: Pentium IV or above
- RAM: 4 GB or more
- Hard Disk: 2GB or more

2.1.2 Software Requirements:

- Operating System: Windows 7 or above
- IDE: Google Colab

2.2 Tools/Languages/Platforms

- Python

Chapter 3 DESIGN AND IMPLIMENTATION

3.1 Architecture/ DFD/Sequence diagram/Class diagrams /Flowchart

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

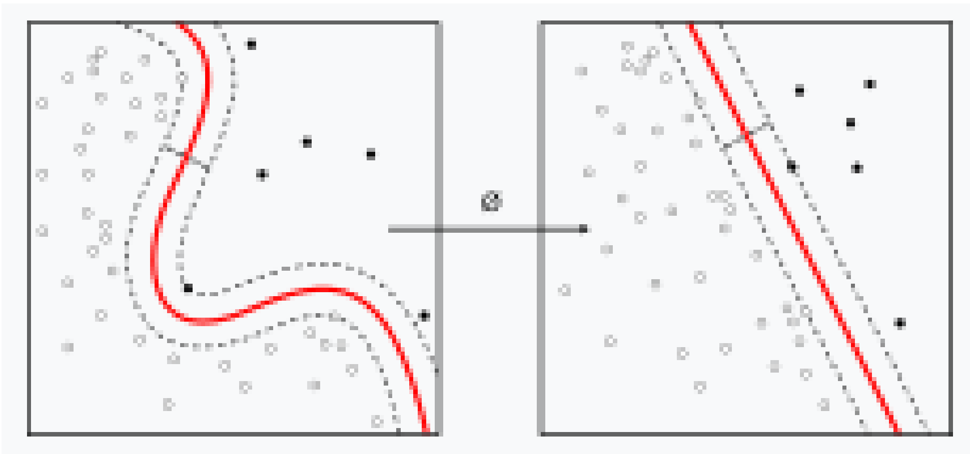


Figure 3.1 Gradient boosting regression Model

3.2 Problem Statement

The goal of this statistical analysis is to help us understand the relationship between insurance prices and how these variables are used to predict insurance price.

Gradient boosting regression Model has been used in terms of minimizing the difference between predicted and actual rating.

The following features have been used:

1. Age: Describes the age of the individual who is applying for the medical insurance policy. The older a person gets, the less healthy he becomes and therefore a higher premium will be required to get a health insurance.
2. Sex: Describes the sex of the individual who is applying for the medical insurance policy. Statistically males are more likely to get into accidents and therefore have a higher premium rates on the insurance policy.
3. BMI: Body mass index (BMI) is a measure of body fat based on height and weight that applies to adult men and women. A good BMI number for the person's height and weight indicates that the person is healthy therefore a lesser premium would be ideal. The normal BMI values can range anywhere between 19.5 and 24.9 for both adult males and females for average height and weight.
4. Children: Indicates the number of children a person has. More the number of children, higher will be the premium as the insurance covers the entire family of the person which includes the spouse and their children.
5. Smoking: Indicates whether a person smokes or not. Smokers generally tend to have higher risk of health issues. Smoking contributes majorly in the hiking up of premiums for health insurance.
6. Region: Indicates the locality in which a person stays. This has a profound effect as some areas are safer to live in due to the lower accident rates, crime rates etcetera. Therefore such safer areas will have lower insurance premiums while the riskier parts of the city will have a higher premium.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Figure 3.2 Description of the Dataset

The above fig3.2, shows the description of the dataset.

3.3 Algorithm

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \underset{\gamma}{\arg\min} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:

   1. Compute so-called *pseudo-residuals*:

   $$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

   2. Fit a base learner (or weak learner, e.g. tree) closed under scaling $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
   3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

   $$\gamma_m = \underset{\gamma}{\arg\min} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

   4. Update the model:

   $$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

3.4 Libraries

- Pandas
- Numpy
- Plotly
- sklearn
- Seaborn
- Matplotlip

Chapter 4

OBSERVATION AND RESULTS

4.1 Testing

Evaluation on Test Data

```
from sklearn.ensemble import GradientBoostingRegressor
gb = GradientBoostingRegressor(random_state=0)
gb.fit(x_train,y_train)
grad_train_pred = gb.predict(x_train)
grad_test_pred = gb.predict(x_test)
GbMae = metrics.mean_absolute_error(y_test, grad_test_pred)
GbMse = metrics.mean_squared_error(y_test, grad_test_pred)
GbRmse = np.sqrt(metrics.mean_squared_error(y_test, grad_test_pred))
GbVar = metrics.explained_variance_score(y_test,grad_test_pred)
##Evaluating the performance of the algorithm
print('Mean Absolute Error:%.2f' %GbMae )
print('Mean Squared Error:%.2f' %GbMse )
print('Root Mean Squared Error:%.2f' %GbRmse )
print('Varscore:%.2f' %GbVar)
```

Visualizing Our predictions based on different algorithms

```
predict = pd.DataFrame(data = models, columns=['Model','MAE', 'MSE', 'RMSE', 'Variance Score'])
```

Perfect predictions

```
new_gb = GradientBoostingRegressor(learning_rate=0.01,n_estimators=400) new_gb.fit(x_train,y_train)
```

4.2 Results & Snapshots

Loading Data

```
In [2]: data = pd.read_csv("medical_cost.csv")

In [3]: data.shape
Out[3]: (1338, 7)

In [4]: data.head()
Out[4]:
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Figure 4.1 Reading CSV File

In the above fig 4.1, we are reading the dataset.csv file and displaying the head.

Data Correlation

In [22]:
```
data.corr()['charges'].sort_values()
```

Out[22]:
```
region     -0.006208
sex         0.057292
children    0.067998
bmi         0.198341
age         0.299008
smoker      0.787251
charges     1.000000
Name: charges, dtype: float64
```

Figure 4.2 Price prediction model

In the above fig4.2, we are visualizing the categories and how much each category affects the price of the insurance policy. With distribution plot of price, we can visualize all the conditions that affect the price and by how much each condition affects the price.

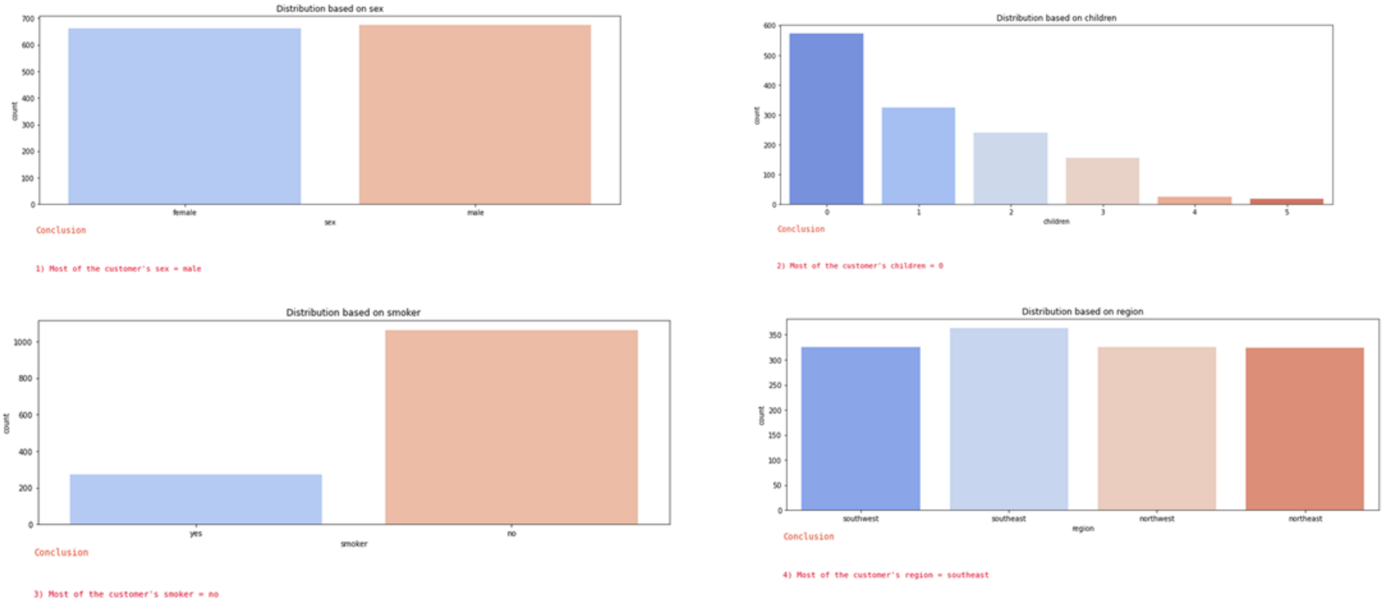
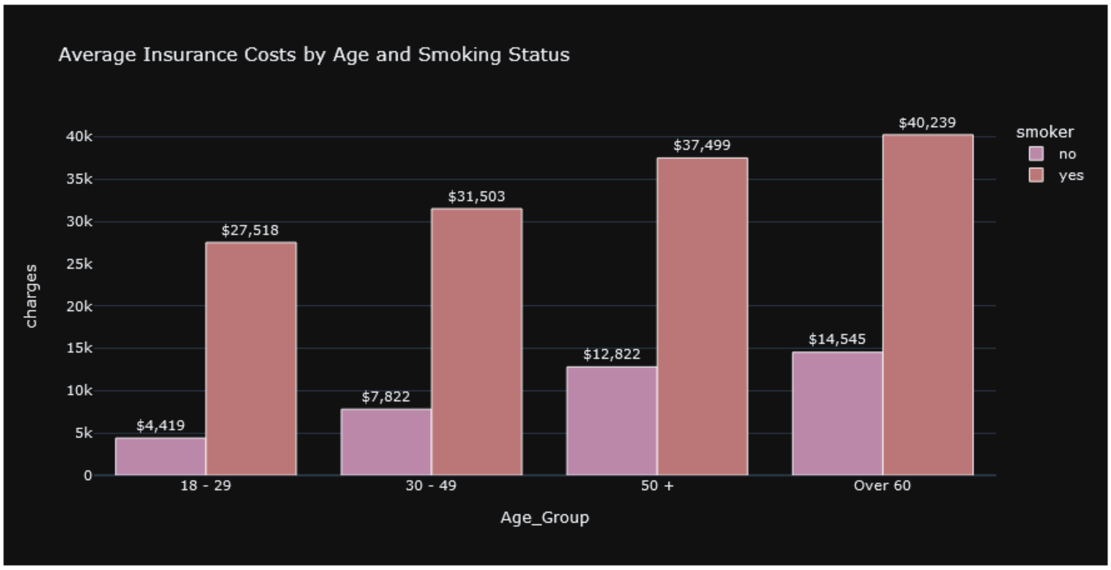


Figure 4.3 Data Analysis

In the above fig4.3, we are visualizing how the data is organized based on sex, number of children, region and if the individual smokes or not. And based on this organization it is determined that majority customers are male, non-smokers with no children living in southwest region. This helps the insurance company to have a standard to compare to.



From the above graph we can say that due to increasing age and If the person is a Smoker than charges are Increasing

Figure 4.4 Price range for smokers vs non-smokers

In the above fig4.4, we are viewing the price change that is happening for people of a certain age group based on whether they smoke or not. Likewise each condition is examined against each other and 4 different algorithms are implemented to predict the price. In this the algorithm with the highest varscore is used in our project.

```
x = data.drop(['charges'], axis = 1)
y = data.charges

x_train,x_test,y_train,y_test = train_test_split(x,y, random_state = 0)
lr = LinearRegression().fit(x_train,y_train)

y_train_pred = lr.predict(x_train)
y_test_pred = lr.predict(x_test)

LrMae = metrics.mean_absolute_error(y_test, y_test_pred)
LrMse = metrics.mean_squared_error(y_test, y_test_pred)
LrRmse = np.sqrt(metrics.mean_squared_error(y_test, y_test_pred))
LrVar = metrics.explained_variance_score(y_test,y_test_pred)

print('Mean Absolute Error: %.2f' %LrMae )
print('Mean Squared Error:%.2f'%LrMse )
print('Root Mean Squared Error:%.2f' %LrRmse)
print('Varscore:%.2f'%LrVar)
```

```
Mean Absolute Error: 3998.27
Mean Squared Error:32073628.56
Root Mean Squared Error:5663.36
Varscore:0.80
```

Figure 4.5 Linear Regression model

In the above fig4.5, we use the linear regression model to predict the prices for insurance and we end up getting an accuracy of 80%.

```
X = data.drop(['charges','region'], axis = 1)
Y = data.charges


quad = PolynomialFeatures (degree = 2)
x_quad = quad.fit_transform(X)

X_train,X_test,Y_train,Y_test = train_test_split(x_quad,Y, random_state = 0)

plr = LinearRegression().fit(X_train,Y_train)

Y_train_pred = plr.predict(X_train)
Y_test_pred = plr.predict(X_test)


PrMae = metrics.mean_absolute_error(Y_test, Y_test_pred)
PrMse = metrics.mean_squared_error(Y_test, Y_test_pred)
PrRmse =  np.sqrt(metrics.mean_squared_error(Y_test, Y_test_pred))
PrVar = metrics.explained_variance_score(Y_test,Y_test_pred)
print('Mean Absolute Error:%.2f' %PrMae)
print('Mean Squared Error:%.2f' %PrMse )
print('Root Mean Squared Error:%.2f'%PrRmse)
print('Varscore:%.2f' %PrVar)
```

```
Mean Absolute Error:2761.13
Mean Squared Error:18117605.54
Root Mean Squared Error:4256.48
Varscore:0.89
```

Figure 4.6 Polynomial Regression model

In the above fig4.5, we use the Polynomial Regression model to predict the prices for insurance and we end up getting an accuracy of 89%.

```
forest = RandomForestRegressor(n_estimators = 100,
                              criterion = 'mse',
                              random_state = 1,
                              n_jobs = -1)
forest.fit(x_train,y_train)
forest_train_pred = forest.predict(x_train)
forest_test_pred = forest.predict(x_test)

RfMae = metrics.mean_absolute_error(y_test, forest_test_pred)
RfMse = metrics.mean_squared_error(y_test, forest_test_pred)
RfRmse = np.sqrt(metrics.mean_squared_error(y_test, forest_test_pred))
RfVar = metrics.explained_variance_score(y_test,forest_test_pred)
##Evaluating the performance of the algorithm
print('Mean Absolute Error:%.2f' %RfMae )
print('Mean Squared Error:%.2f' %RfMse )
print('Root Mean Squared Error:%.2f' %RfRmse )
print('Varscore:%.2f' %RfVar)
```

```
Mean Absolute Error:2705.78
Mean Squared Error:19965476.41
Root Mean Squared Error:4468.27
Varscore:0.88
```

Figure 4.7 Random Forest Regression model

In the above fig4.7, we use the Random Forest regression model to predict the prices for insurance and we end up getting an accuracy of 88%.

```
from sklearn.ensemble import GradientBoostingRegressor
gb = GradientBoostingRegressor(random_state=0)
gb.fit(x_train,y_train)

grad_train_pred = gb.predict(x_train)
grad_test_pred = gb.predict(x_test)

GbMae = metrics.mean_absolute_error(y_test, grad_test_pred)
GbMse = metrics.mean_squared_error(y_test, grad_test_pred)
GbRmse = np.sqrt(metrics.mean_squared_error(y_test, grad_test_pred))
GbVar = metrics.explained_variance_score(y_test,grad_test_pred)
##Evaluating the performance of the algorithm
print('Mean Absolute Error:%.2f' %GbMae )
print('Mean Squared Error:%.2f' %GbMse )
print('Root Mean Squared Error:%.2f' %GbRmse )
print('Varscore:%.2f' %GbVar)
```

```
Mean Absolute Error:2433.88
Mean Squared Error:16012531.39
Root Mean Squared Error:4001.57
Varscore:0.90
```

Figure 4.8 Gradient boosting regression model

In the above fig4.7, we use the Gradient boosting regression model to predict the prices for insurance and we end up getting an accuracy of 90%.

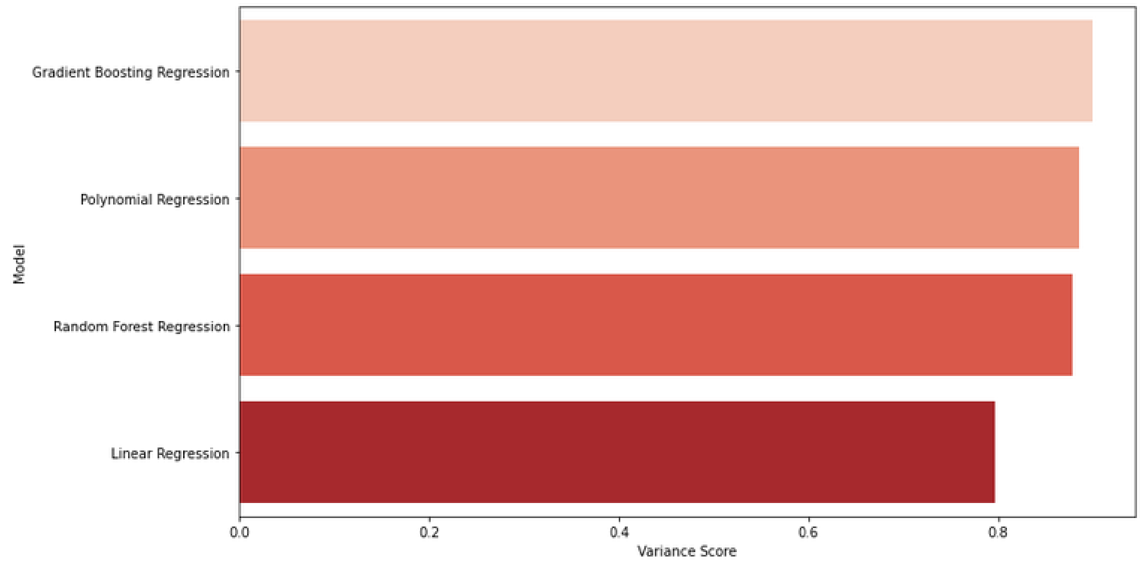


Figure 4.9 Visualizing Accuracies of different algorithms

In the above fig4.9, we compare the algorithms that predict the price of insurance based on the different conditions provided. It is evident that the gradient booster regression algorithm has the highest accuracy of 90%. Since this model provides the highest accuracy for price prediction, we use this algorithm to determine the final output.

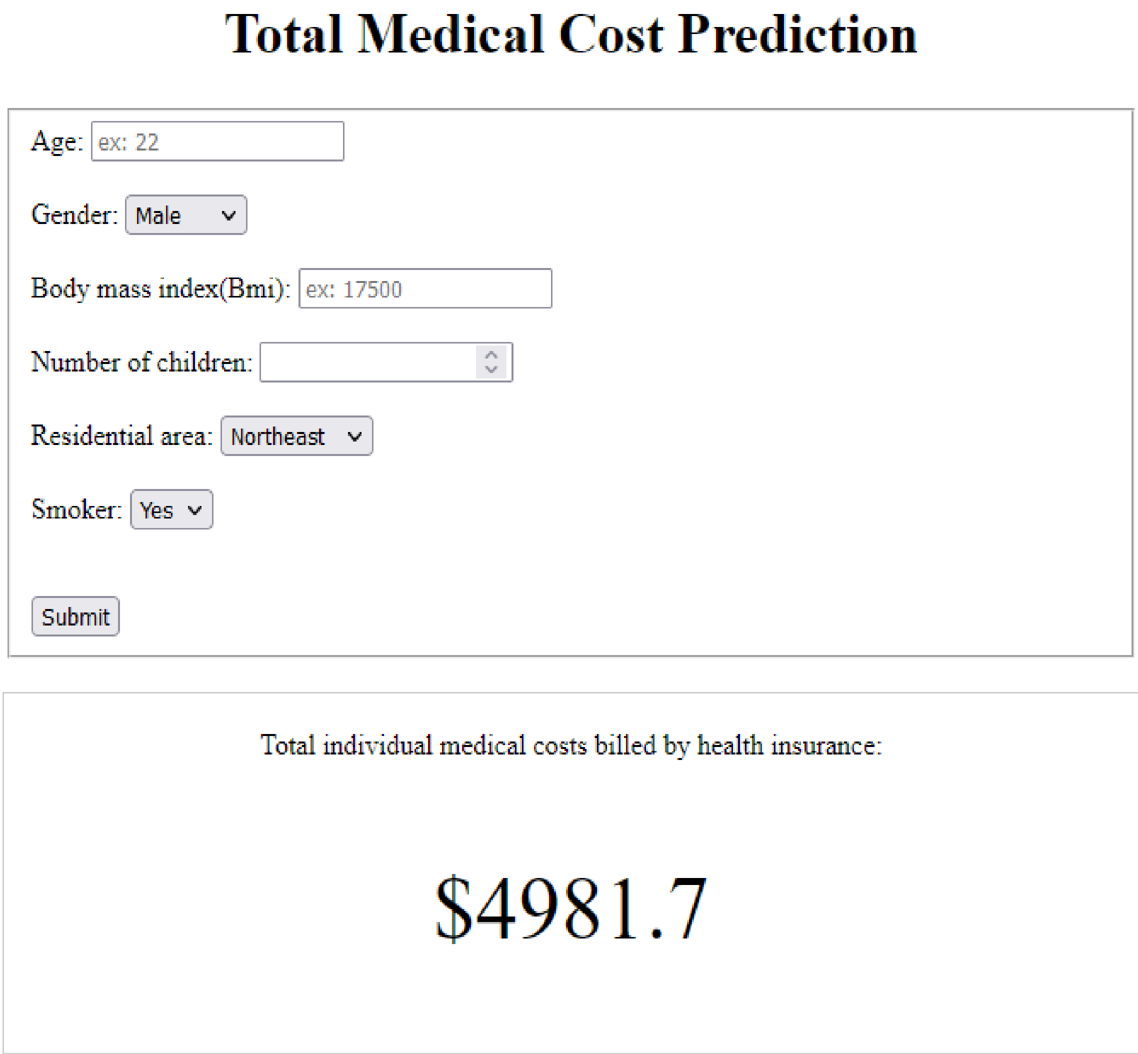


Figure 4.9.1 Result

In the above fig4.9.1, we are calculating the actual cost for a person to get health insurance based on his age, bmi, no.of children, area, sex and smoking status. For the values of 22, 17500, 0, southeast, male and yes, we get an estimated price of 4981.7$ to get health insurance for that person. This value has an accuracy of 90% being the final insurance value.

Chapter 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

The proposed model is the best substitute for the manual method where a third party is involved as the middleman and is potentially vulnerable along with it being cheaper for the end customers. Based on the results, it can be concluded that such ML-driven predictions are easily comprehendible and significant from a data-analytics point of view.
When correctly implemented, a high rate of accuracy can be achieved.

5.2 Future Enhancement

- To make the interface more informative and user-friendly by implementing better GUI designs.
- Using bigger training data sets to get a more accurate estimate of the prices.
- Implementing other machine learning algorithms which can improve the accuracy of the model.

Chapter 6

REFERENCE

1) Multi-View Deep Learning Framework for Predicting Patient Expenditure in Healthcare XIANLONG ZENG 1, SIMON LIN2, AND CHANG LIU 3 (Member, IEEE)
DOI: 10.1109/OJCS.2021.3052518

2) Medicine Expenditure Prediction via a VarianceBased Generative Adversarial Network SHRUTI KAUSHIK 1 , (Member, IEEE), ABHINAV CHOUDHURY1 , SAYEE NATARAJAN2 , LARRY A. PICKETT, JR.2 , AND VARUN DUTT1 , (Senior Member, IEEE) **DOI:** 10.1109/ACCESS.2020.3002346

3) Medicine Expenditure Prediction via a Variance- Based Generative Adversarial Network by Abhinav Choudhury; Sayee Natarajan; Larry A. Pickett; Varun Dutt
DOI: 10.1109/ACCESS.2020.3002346