Mini-Project – 2B Web based on ML (ITM 601)

Medical Insurance Cost Recommendation

T. E. Information Technology

By

Harshkumar Bhikadiya	**29**
Shivam Bhosale	**30**
Parth Dali	**31**
Pranav Dalvi	**32**

Mentor:

Dr. Nitika Rai
Associate Professor

Department of Information Technology
St. Francis Institute of Technology
(Engineering College)
University of Mumbai
2021-2022

CERTIFICATE

This is to certify that the project entitled “**Medical Insurance Cost Recommendation**” is a bonafide work of **Harsh Bhikadiya (29), Shivam Bhosale (30), Parth Dali (31) and Pranav Dalvi (32)** submitted to the University of Mumbai towards completion of mini project work for the subject of **Mini-Project -2B Web based on ML (Course Code: ITM601).**

Dr. Nitika Rai
Supervisor/Guide

Dr. Joanne Gomes
HOD-IT

Examiners

1.--

2.--

Date:

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

------------------------------------- -------------------------------------

(Name of student and Signature) (Name of student and Signature)

------------------------------------- -------------------------------------

(Name of student and Signature) (Name of student and Signature)

ABSTRACT

Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. This study demonstrates how linear regression can forecast insurance costs.

Dataset was used for training the models and that training helped to come up with some predictions. Then the predicted amount was compared with the actual data to test and verify the model. Later the accuracy of the model was evaluated.

Linear regression was used as it is simple to implement and easier to interpret the output coefficients.

Index

List of Abbreviations

Sr. No.	Abbreviation	Full Form
1	INR	Indian Rupee

List of Figures

Chapter 1
Introduction

1.1 Background:

With the constantly increasing prices of healthcare in our country, and with the ever-rising instances of diseases, health insurance today is a necessity.

Health insurance provides people with a much-needed financial backup at times of medical emergencies.

1.2 Scope of the project:

As more and more healthcare and medical companies are witnessing the value of AI and machine learning within their varied systems, industry leaders are realizing that machine learning applications can potentially improve the accuracy of treatment protocols and health outcomes.It’s important to note that a majority of the price consumers pay when enrolling in health insurance goes into risk prediction and risk management. By using AI to create a system that can create more accurate risk models and predict which individuals need specific types of care, health insurance providers can spend more money on their beneficiaries and less on those processes.

1.3 Objectives and Problem Statement:

Objectives

1.To build a Machine learning model to recommend insurance policy.
2.To help the customer to be financially prepared in case of medical emergency

Problem Statement

To develop a machine learning model which can recommend the cost of insurance policy which a customer should purchase. This model will be able to recommend the cost of insurance to the customers based on their BMI, age, medical history, and smoking habits.

Chapter 2
Literature Review

In this section, research efforts from the exploration of information and machine learning techniques are discussed. Several papers have discussed the issue of claim prediction.

Jessica Pesantez-Narvaez suggested, "Predicting motor insurance claims using telematics data" in 2019. This research compared the performance of logistic regression and XGBoost techniques to forecast the presence of accident claims by a small number and results showed that because of its interpretability and strong predictability [3], logistic regression is an effective model than XGBoost.

System proposed by Ranjodh Singh et al in 2019, this system takes pictures of the damaged car as inputs and produces relevant details, such as costs of repair to decide on the amount of insurance claim and locations of damage. Thus the predicted car insurance claim was not taken into account in the present analysis but was focussed on calculating repair costs [4].

Oskar Sucki 2019, The purpose of this research is to study the prediction of churn. Random forests were considered to be the best model (74 percent accuracy). In some fields, the dataset had missing values. Following an analysis of the distributions, the decision has been taken to substitute the missing variables with additional attributes suggesting that this data does not exist [5]. This is permitted only if the data is absolutely randomly lost, and so the missing data mechanism by which the appropriate approach to data processing is decided has first to be established [6][7].

In 2018, Muhammad Arief Fauzan et al. In this paper, the exactness of XGBoost is applied to predict statements. Compare the output with the performance of XGBoost, a collection of techniques e.g., AdaBoost, Random Forest, Neural Network. XGBoost offers better Gini structured accuracy. Using publicly accessible Porto Seguro to Kaggle datasets. The dataset includes huge quantities of NaN values but this paper manages missing values by medium and median replacement. However, these simple, unprincipled methods have also proven to be biased [7]. They, therefore, concentrate on exploring the methods ML that are highly appropriate for the problems of several missing values, such as XGboost[8].

G. Kowshalya, M. Nandhini. in 2018. Three classifiers have been developed in this study to predict and estimate fraudulent claims and a percentage of premiums for the various customers based upon their personal and financial data. For classification, the algorithms Random Forest, J48, and Naïve Bayes are chosen. The findings show that Random Forest exceeds the remaining techniques depending on the synthetic dataset. This paper therefore does not cover insurance claim forecasts, but rather focuses on false claims [9]. The above previous works did not consider both predicted the cost or claim severity, they only make a classification for the issues of claims (whether or not a claim was filed for that policyholder) in this study we focus on advanced statistical methods and machine learning algorithms and deep neural network for predict the cost of health insurance.

Chapter 3
Proposed Work

3.1 Architectural Details

Fig 3.1.1 shows the architectural diagram which depicts the steps to be followed.

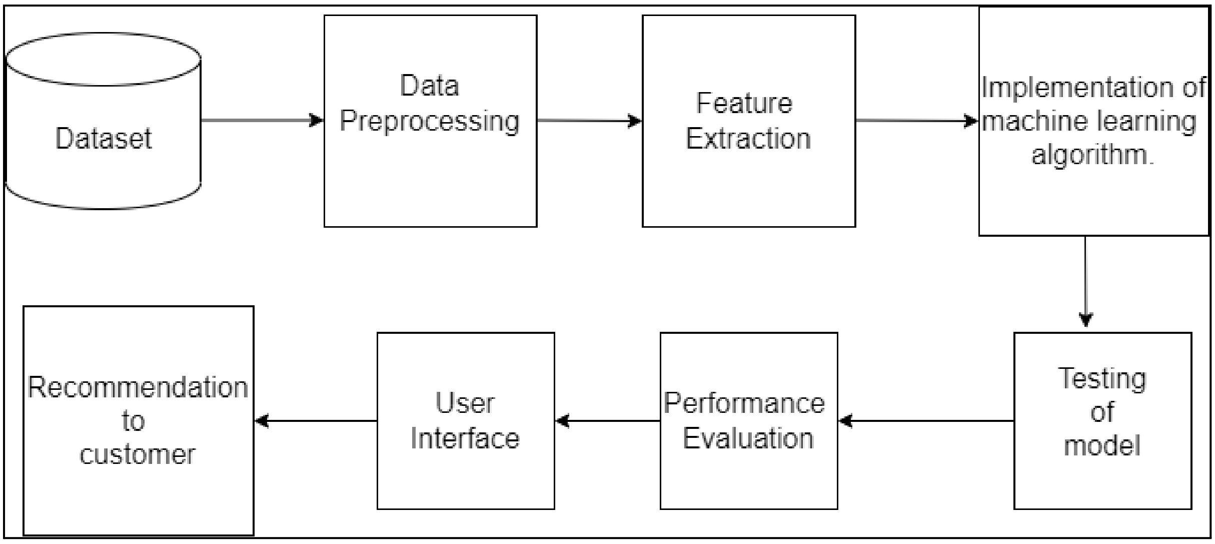


Fig. 3.1.1 Architectural Diagram

3.1.1 Data Collection

Data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

Data Collection Techniques:

- Interviews.
- Questionnaires.
- Financial Statements
- Sales Reports
- Retailer/Distributor/Deal Feedback
- Customer Personal Information (e.g., name, address, age, contact info)
- Business Journals
- Government Records (e.g., census, tax records, Social Security info)
- Trade/Business Magazines
- The internet

3.1.2 Feature Engineering

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

Feature engineering consists of various process:

- Feature Creation
- Transformations
- Feature Extraction
- Exploratory Data Analysis
- Benchmark

3.1.3 Implementation of machine learning algorithm

Implementing a machine learning algorithm will give you a deep and practical appreciation for how the algorithm works. This knowledge can also help you to internalize the mathematical description of the algorithm by thinking of the vectors and matrices as arrays and the computational intuitions for the transformations on those structures.

Following factors should be taken into account while choosing an algorithm:

- The kind of model in use (problem)
- Analyzing the available Data (size of training set)
- The accuracy of the model
- Time taken to train the model (training time)
- Number of parameters
- Number of features
- Linearity

3.1.4 Testing of model

In machine learning, model testing is referred to as the process where the performance of a fully trained model is evaluated on a testing set.Training dataset is independent of the training set but has a somewhat similar type of probability distribution of classes and is used as a benchmark to evaluate the model, used only after the training of the model is complete. Testing set is usually a properly organized dataset having all kinds of data for scenarios that the model would probably be facing when used in the real world.This data is approximately 20-25% of the total data available for the project.

3.1.5 Performance Evaluation

Performance evaluation is an important aspect of the machine learning process. However, it is a complex task. It, therefore, needs to be conducted carefully in order for the application of machine learning

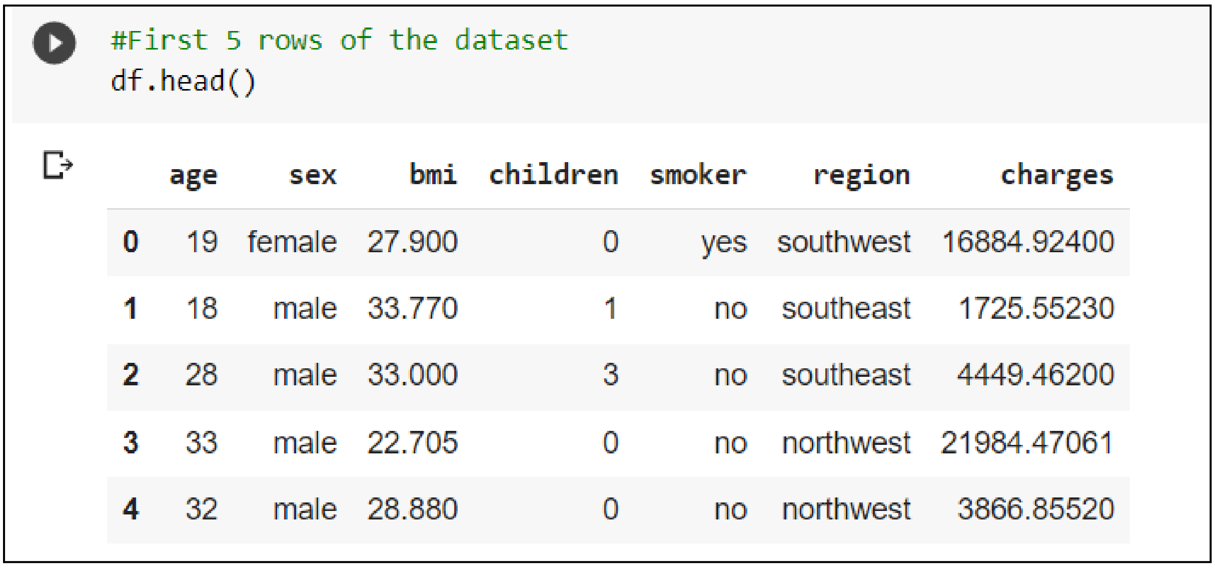
The various ways to check the performance of our machine learning model are:

- Confusion matrix
- Accuracy
- Precision
- Recall
- F1 score
- Precision-Recall or PR curve
- ROC (Receiver Operating Characteristics) curve
- PR vs ROC curve.

Chapter 4
Implementation

4.1 Dataset Details

Fig. 4.1 shows the fields present in the dataset.



```
#First 5 rows of the dataset
df.head()
```

	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

Fig. 4.1 Data set details

In data collection we used the Kaggle dataset. The attributes of the dataset are:

- age: age of the 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/m²) 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: smoking or not
- region:the beneficiary's residential area in the northeast, southeast, southwest, northwest.
- charges: individual medical costs billed by health insurance.

4.2 Algorithm details

Linear Regression Algorithm:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the image shown in Fig. 4.2.

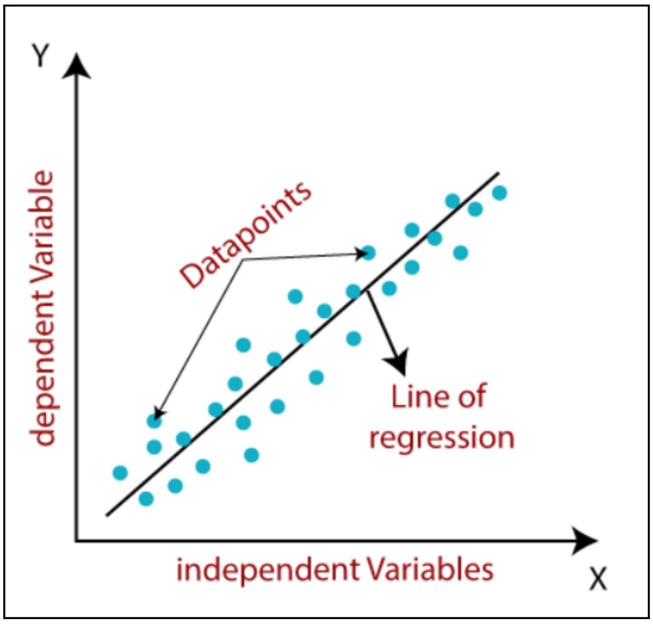


Fig. 4.2 Linear regression

Mathematically, we can represent a linear regression as:

y= a0+a1x+ ε

Here,

y= Dependent Variable (Target Variable)
x= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

4.3 Performance metrics in detail

R-Squared value

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

Interpretation of R-Squared

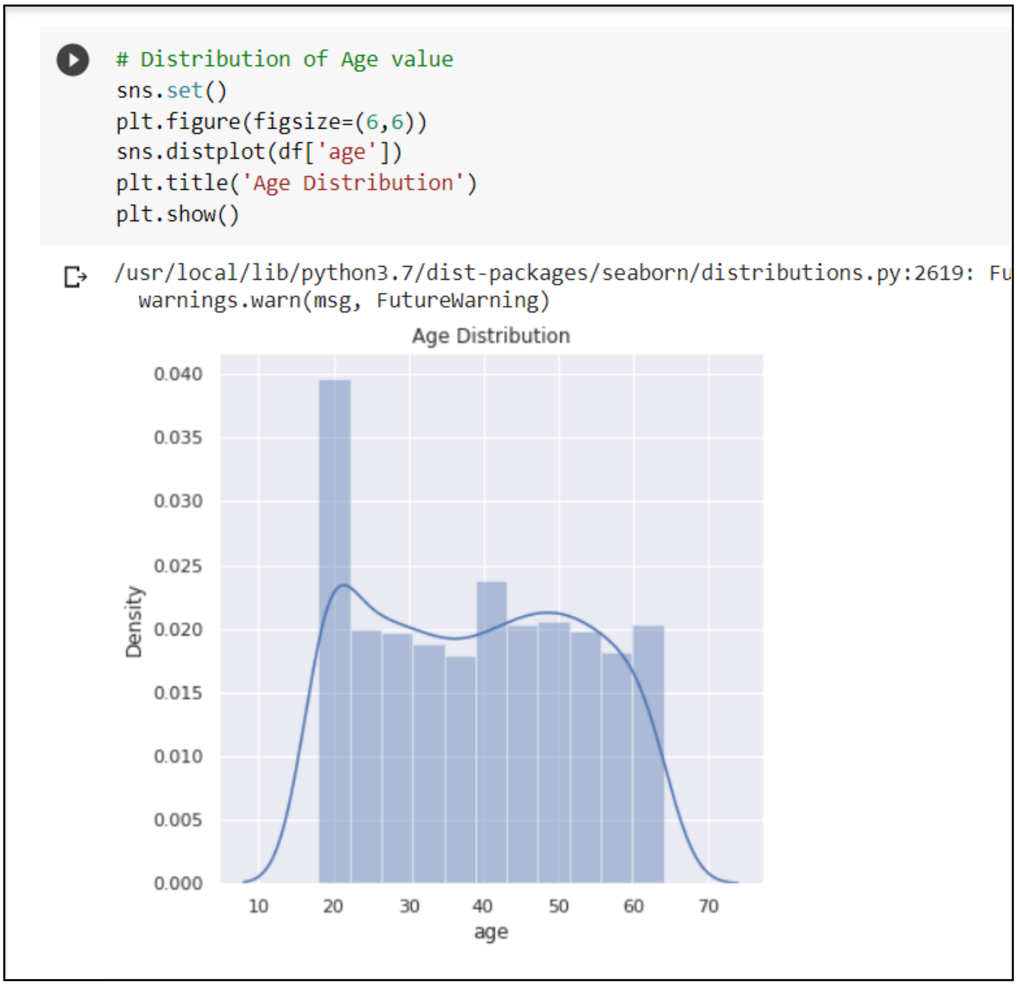
The most common interpretation of r-squared is how well the regression model fits the observed data. For example, an r-squared of 60% reveals that 60% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

Chapter 5
Results and Discussions

5.1 Discussions:

5.1.1: Age Distribution

From the Figure 5.1.1, it was observed that the average age of the customers is around 20-30 years old.



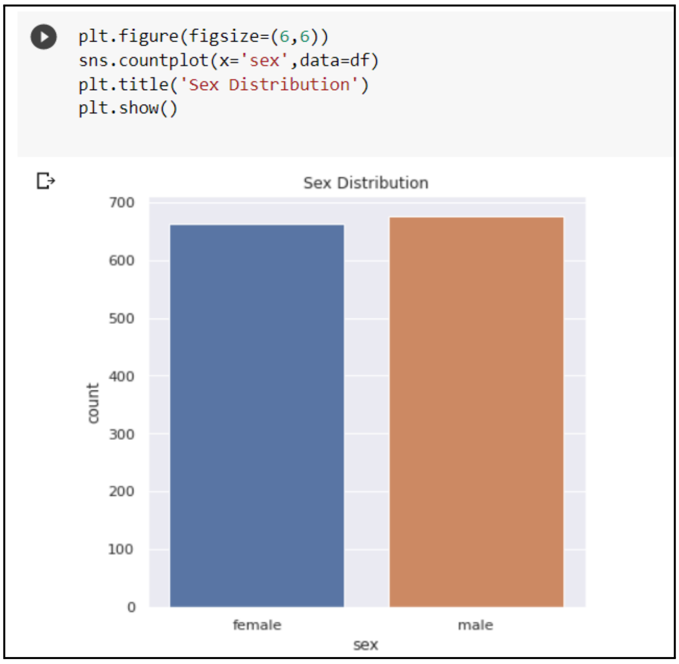
```
# Distribution of Age value
sns.set()
plt.figure(figsize=(6,6))
sns.distplot(df['age'])
plt.title('Age Distribution')
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: Fu
  warnings.warn(msg, FutureWarning)
```

Fig. 5.1.1 Age Distribution

5.1.2: Sex Distribution

From figure 5.2.2, It was observed that male and female are in equal ratio in the dataset.

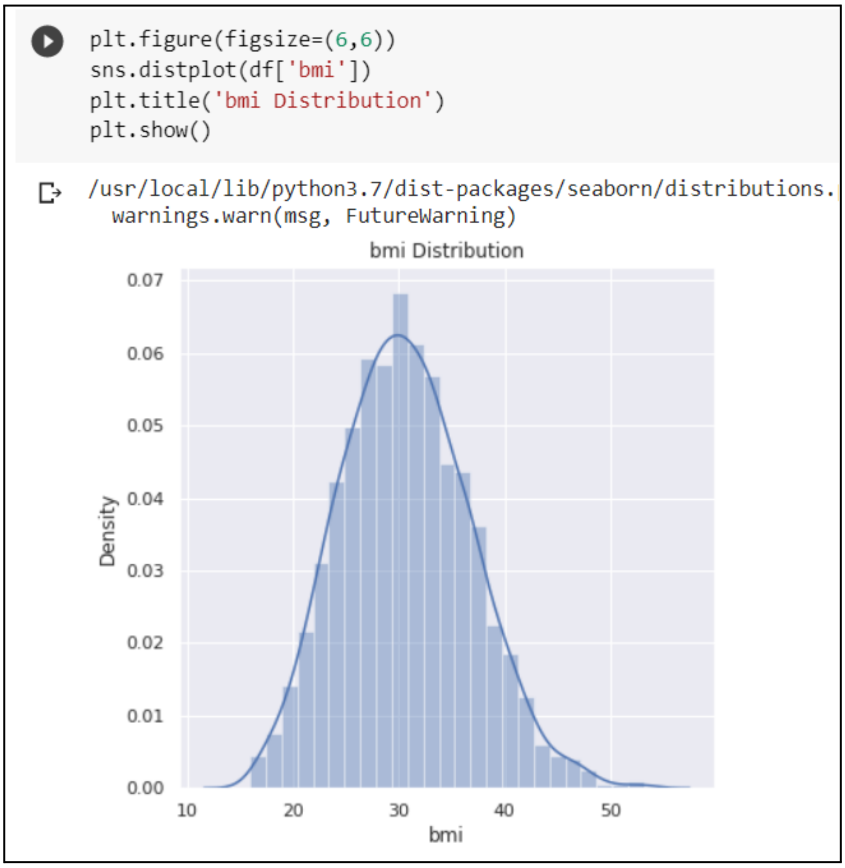


```
plt.figure(figsize=(6,6))
sns.countplot(x='sex',data=df)
plt.title('Sex Distribution')
plt.show()
```

Fig. 5.1.2 Sex Distribution

5.1.3: BMI Distribution

From figure 5.1.3, The average BMI observed was between 30-35.



```
plt.figure(figsize=(6,6))
sns.distplot(df['bmi'])
plt.title('bmi Distribution')
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.
  warnings.warn(msg, FutureWarning)
```

Fig. 5.1.3 BMI Distribution

5.2 Results:

Fig. 5.2.1 shows the Home page redirect to the recommend form

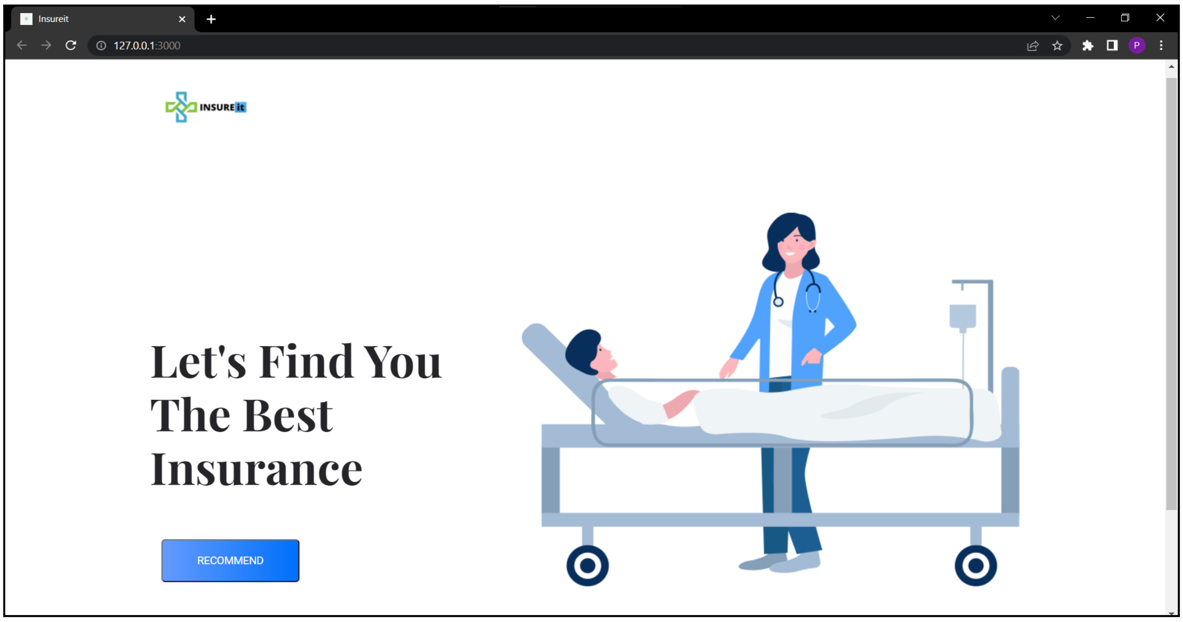
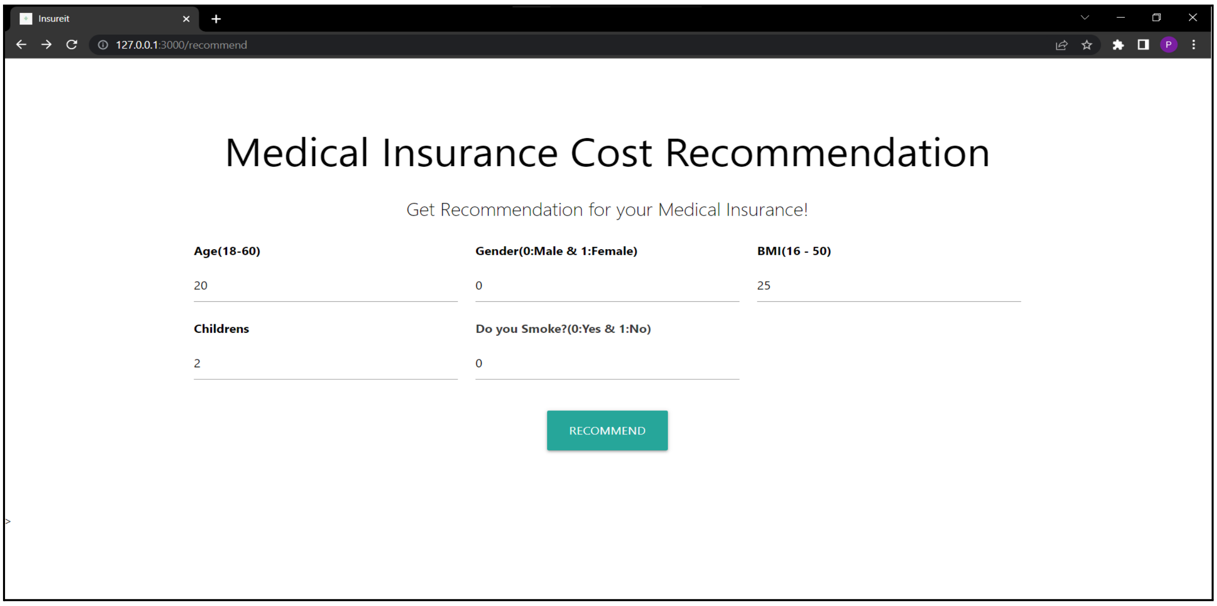


Fig. 5.2.1 Home page

Fig. 5.2.2 shows the page where user enters his medical detail and submit the form



Medical Insurance Cost Recommendation

Get Recommendation for your Medical Insurance!

Age(18-60)
20

Gender(0:Male & 1:Female)
0

BMI(16 - 50)
25

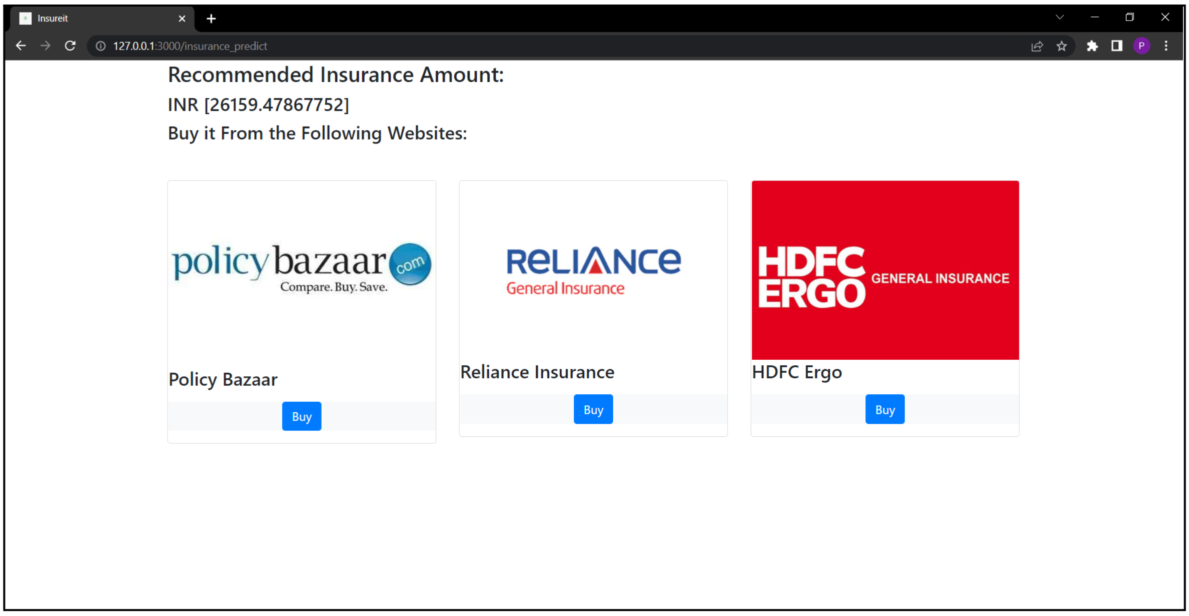
Childrens
2

Do you Smoke?(0:Yes & 1:No)
0

RECOMMEND

Fig. 5.2.2 Input data

Fig. 5.2.3 displays the recommended amount and recommends the websites to buy the policies.



Recommended Insurance Amount:

INR [26159.47867752]

Buy it From the Following Websites:

Fig. 5.2.3 Final page

Chapter 6
Conclusion and Future Scope

6.1 Conclusion:

In this project, we used a linear regression algorithm for evaluating individual health insurance data. The predicted premiums from this model were compared with actual premiums to compare the accuracy of the model.

Various factors were used and their effect on the predicted amount was examined. It was observed that a person’s age and smoking status affects the prediction most.

Premium amount prediction focuses on a person's own health rather than other company’s insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance.

6.2 Future Scope:

The current machine learning model is limited to accurate predictions of the specific age group hence in future scope we can increase the quality and accuracy of the dataset. The user interface can be improved and buying of new policies can be incorporated.

Inclusion of more parameters like chronic diseases, number of surgeries can be used in the future for better cost recommendation.

References

[1] Mohamed hanafy, Omar M. A. Mahmoud."Predict Health Insurance Cost by using Machine Learning and DNN Regression Models" International Journal of Innovative Technology and Exploring Engineering (IJITEE)(2021):2278-3075

[2] Kaggle[online].Available https://www.kaggle.com/datasets/awaiskaggler/insurance-csv(Accessed:Jan 23, 2022)

[3] Researchgate.net.[Online].Available https://www.researchgate.net/publication/348559741_Predict_Health_Insurance_Cost_by_using_Machine_Learning_and_DNN_Regression_Models. (Accessed: 24-Jan-2022).

[4] moneycrashers.[online].Available https://www.moneycrashers.com/factors-health-insurance-premium- costs(Accessed:Jan 23, 2022)

Acknowledgment

We are thankful to our college **St. Francis Institute of Technology** for giving us this chance to gain exposure in solving real world problems and acquire practical knowledge and skill sets that will prove to be very crucial to our long term career prospects. We would take this opportunity to express our sincerest gratitude to our esteemed director **Bro. Jose Thuruthiyil**, our principal **Dr. Sincy George** and our **HOD, Dr. Joanne Gomes** for their encouragement, the direction that they give to our college and us students, and the facilities provided to us.

This project, and the research that we undertook, could not have been realized without the utmost support of our project guide **Dr. Nitika Rai**, who guided us every step of the way, starting from the conception of the project, right up to the execution of the finished solution.