

A Project Report on

Estimation and Prediction of Hospitalization and Medical Care Costs

by

Team Leader :SHAIK FARIDA(20AT1A0534)
Student 1 :E.LIKHITHA(20AT1A0571)
Student 2 :D.PRANATHI KETURA(20AT1A05A5)
Student 3 :M.SAI SAHITHI(20AT1A05C5)

Under the Guidance of

MRS.M.JAYA SUNITHA,M.Tech
Associate Professor



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

G. PULLAIAH COLLEGE OF ENGINEERING AND TECHNOLOGY
(Autonomous)

(Approved by AICTE | NAAC Accreditation with 'A' Grade | Accredited by NBA (ECE, CSE, EEE, CE) |
Permanently Affiliated to JNTUA)

ABSTRACT

Medical costs are one of the most common recurring expenses in a person's life. Based on different research studies, BMI, ageing, smoking, and other factors are all related to greater personal medical care costs. The estimates of the expenditures of health care related to obesity are needed to help create cost-effective obesity prevention strategies. Obesity prevention at a young age is a top concern in global health, clinical practice, and public health. To avoid these restrictions, genetic variants are employed as instrumental variables in this research.

Using statistics from public huge datasets, the impact of body mass index (BMI) on overall healthcare expenses is predicted. A multiview learning architecture can be used to leverage BMI information in records, including diagnostic texts, diagnostic IDs, and patient traits. A hierarchy perception structure was suggested to choose significant words, health checks, and diagnoses for training phase informative data representations, because various words, diagnoses, and previous health care have varying significance for expense calculation.

In this system model, linear regression analysis, naive Bayes classifier, and random forest algorithms were compared using a business analytic method that applied statistical and machine-learning approaches. According to the results of our forecasting method, linear regression has the maximum accuracy of 97.89 percent in forecasting overall healthcare costs. In terms of financial statistics, our methodology provides a predictive method.

Contents

Abstract	2
Contents	3
List of Figures and Tables	4
CHAPTER 1 INTRODUCTION	5
CHAPTER 2 RELATED WORKS	6
CHAPTER 3 PROPOSED METHOD	7
CHAPTER 4 EXPERIMENTAL RESULTS	8-11
CHAPTER 5 FEATURES/ADVANTAGES	12
CHAPTER 6 RESULTS AND ANALYSIS	13
CHAPTER 6 CONCLUSIONS	14

LIST OF FIGURES AND TABLES

DataBase collected from IBM cognos about the Estimation and prediction of hospitalization and medical care costs

CHAPTER 1 INTRODUCTION

The incidence of overweight and obesity has increased significantly in most countries in recent decades. Excess weight is associated with an increased incidence of many chronic diseases, including vascular disease, respiratory disease, osteoarthritis, some cancer, type 2 diabetes, and premature death. There is consistent evidence that an increased BMI is associated with higher health costs, and these costs are expected to increase as obesity. Modelling uses machine-learning methods, in which the machine learns from the data and uses it to forecast new data. The most commonly predictive analytic model used is regression . The proposed model for accurate prediction of future outputs has applications in banking, economics, e-commerce, sports, business, entertainment, etc. A method used to forecast healthcare costs for BMI is based on several factors. Multiple linear regression is one of the statistical techniques for estimating the relationship among the dependent (target) and independent variables. The regression method is commonly used to develop a system based on a number of factors to predict the cost .

The regression analysis is performed to determine the relationship among two or more variables with cause-effect relationships and to make predictions for the topic using the relationships . If regression used one independent variable, then it is known as univariate regression analysis, or else if it used more than two independent variables then it is known as multivariate regression analysis. Linear regression involves initially uploading the data and then analysing the data. Subsequently, the data are cut, and then, the data are trained and separated to create the model. At last, it will evaluate the accuracy. The main aim of regression is to develop an efficient technique for predicting dependent properties from a set of characteristic variables. A regression problem is the actual or continuous value of the output variables, that is, area, salary, and weight. Regression can be defined as a statistical method used in applications such as predicting the healthcare costs. Regression is used to predict the relationship among the

dependent variable and set of independent variables. There are various types of regression techniques available namely simple linear regression, multiple linear regression, polynomial regression, support vector regression, and random forest regression.

Fast-growing healthcare costs have become a significant challenge in several developed countries. Existing evidence suggests that healthcare costs have accumulated among a large number of BMI. Even though experiments have attempted to develop accurate models for predicting healthcare costs for BMI, their effectiveness is excellent due to the lack of detailed clinical information in the data used to create complex intervals and prognostic models. Numerous studies on more costs for obesity patient prognostic models have relied on self-report data and electronic health data from claims. Data from laboratory tests are defined—these, more granular and detailed clinical information, lead to improvements in the prognostic model. A recent survey by health research program and claim data shows that there is an improvement in the performance of the machine-learning-based predictive model for health costs for obesity. Still, many insurers and providers worldwide are actively seeking an approach that can accurately predict obesity BMI.

However, despite the potential value of advanced machine-learning approaches for risk prediction, payers and providers still rely heavily on linear regression to manage and adapt their patient population . The slow adoption of advanced machine-learning techniques may be partly explained by the lack of familiarity with risk stabilization analysts with such techniques and the combination of complex interpretation and results required in practice. Machine-learning regression models are within the framework of standard linear regression and perform some sophisticated but less explicit machine-learning techniques . This study focused on fine linear regression models, which conducted a complete comparison of penalty regression with linear regression in forecasting overall health costs, which was not reported in the previously published literature. The major focus of this study is to estimate the health costs incurred due to obesity in the population.

The rest of this study is formalized as follows: Section [2](#) defines the related works on estimating the healthcare costs using various methodology methods. Section [3](#) designates in detail the workflow of the proposed algorithm. Section [4](#) represents the experiments with results and comparison graphics with existing works and its discussion. Finally, Section [6](#) concludes the study.

CHAPTER 2

2. Related Work

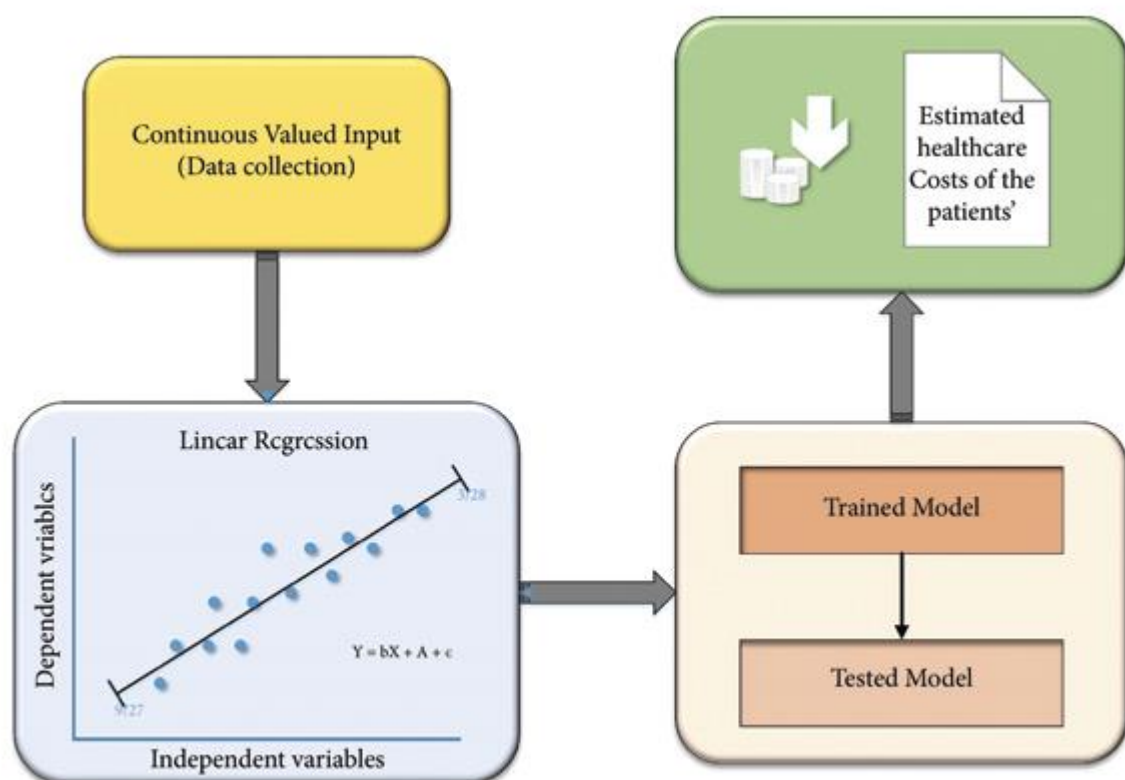
Some of the recent literature that describes the various mechanism of estimating the costs of physical healthcare is summarized below. In unplanned 30-day readmissions are a common occurrence among congestive heart failure (CHF) patients, posing major health concerns and increasing healthcare costs. It is critical to implement tailored treatment programs for high hazard patients of readmission in an attempt to prevent readmissions and lower healthcare costs. This necessitates recognizing high individuals at the time of hospital release. They constructed and evaluated a deep learning network to predict 30-day unplanned readmission using actual information from over 7,500 CHF patients hospitalized in Sweden. Using specialist characteristics and situational integration of medical knowledge provides a cost-sensitive implementation of the long short-term memory (LSTM) neural net. Using both machine-derived and professional characteristics, including frequent patterns, and resolving the issue of class imbalances, this research focuses on important parts of an EHR-driven forecasting system in a single framework. We assess each element's impact on forecasting effectiveness (F1 measure, ROC-AUC) and price benefits. In at least 2 evaluating criteria, it shows that the technique with all critical features outperforms the simplified approaches in terms of discriminating capability. Researchers also propose a basic economic assessment to predict annual income if high-risk patients are provided tailored therapies.

Patients with heart failure (HF) require precise hazard classification to implement tailored therapies focused on enhancing their efficiency of living and results. To assess the economic benefit of complementing claim-based forecasting analytics with electronic medical record (EMR)-derived data and to contrast machine-learning techniques to conventional logistic regression in forecasting critical results in patients with HF, healthcare patients with HF from 2 healthcare professional systems in Massachusetts, Boston, were included in predictive research with a one-year follow-up duration. "Providers" comprise therapists, various medical professionals, clinicians, and their organization including the network. Logistic regression, gradient boosted modelling, regression trees, random forests, least absolute shrinkage, classification, and selection operation regression were used to predict all-cause morbidity, top cost decile, HF hospitalization, gradient boosted modelling, and home days loss larger than 25%. Information from network 1 was used to educate all algorithms, which were then evaluated in network 2. The area under high accuracy curves (AUPRCs) and overall value estimations from decision curves were obtained after choosing the best effective modelling strategy depending on the Brier score, calibration, and discrimination.

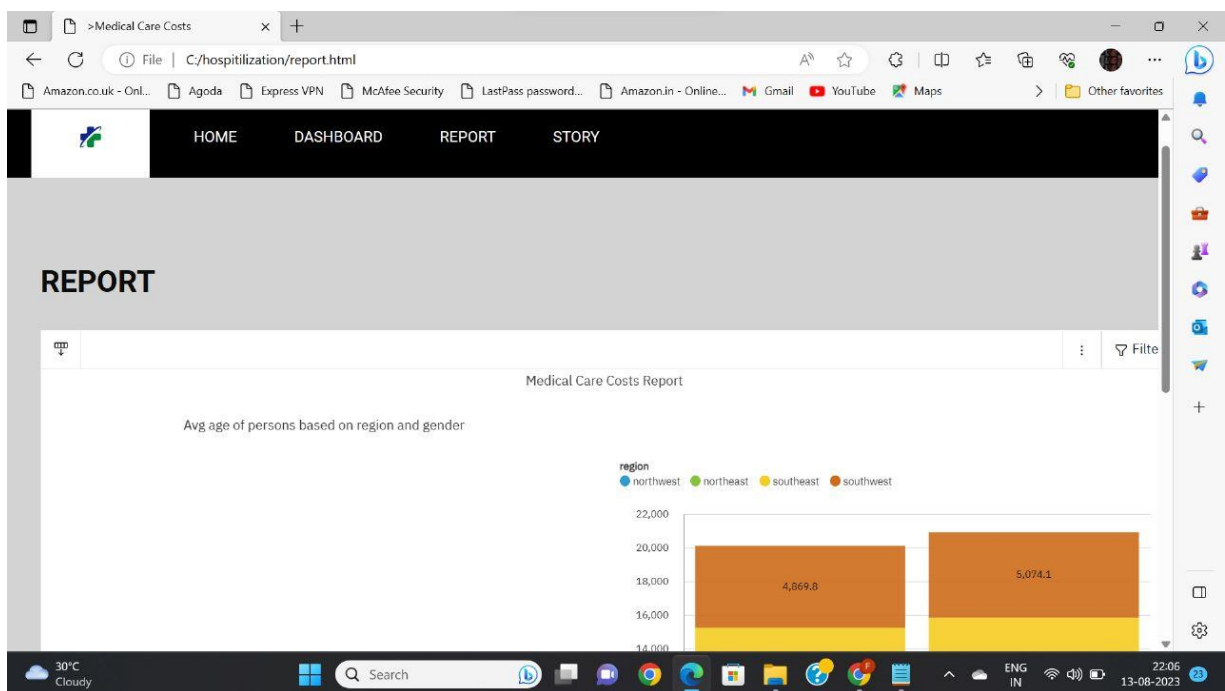
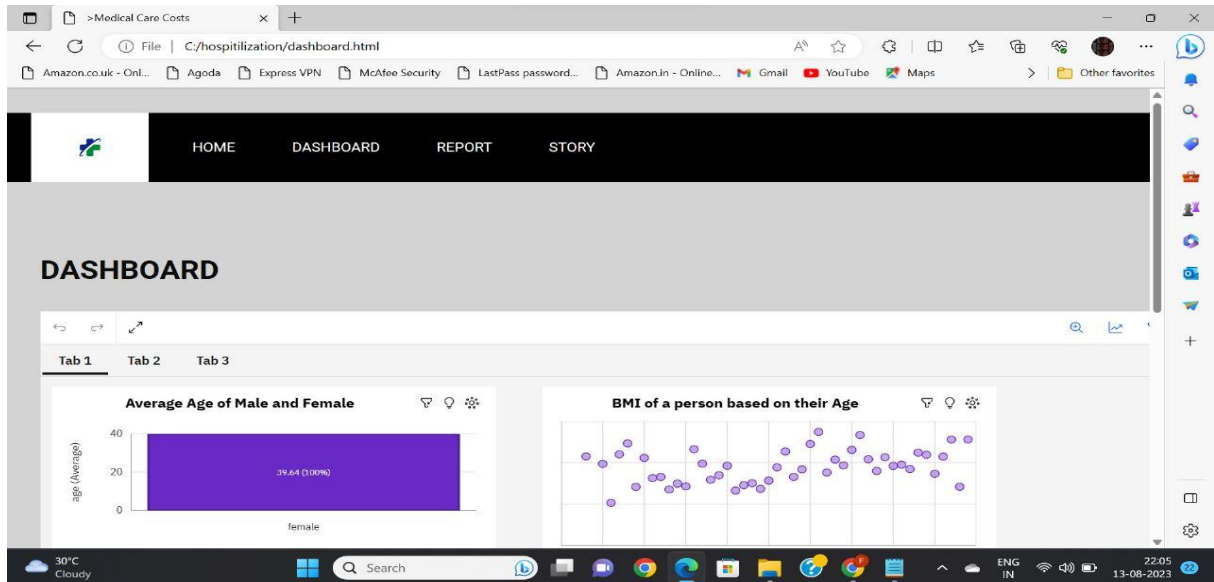
CHAPTER 3 PROPOSED METHOD

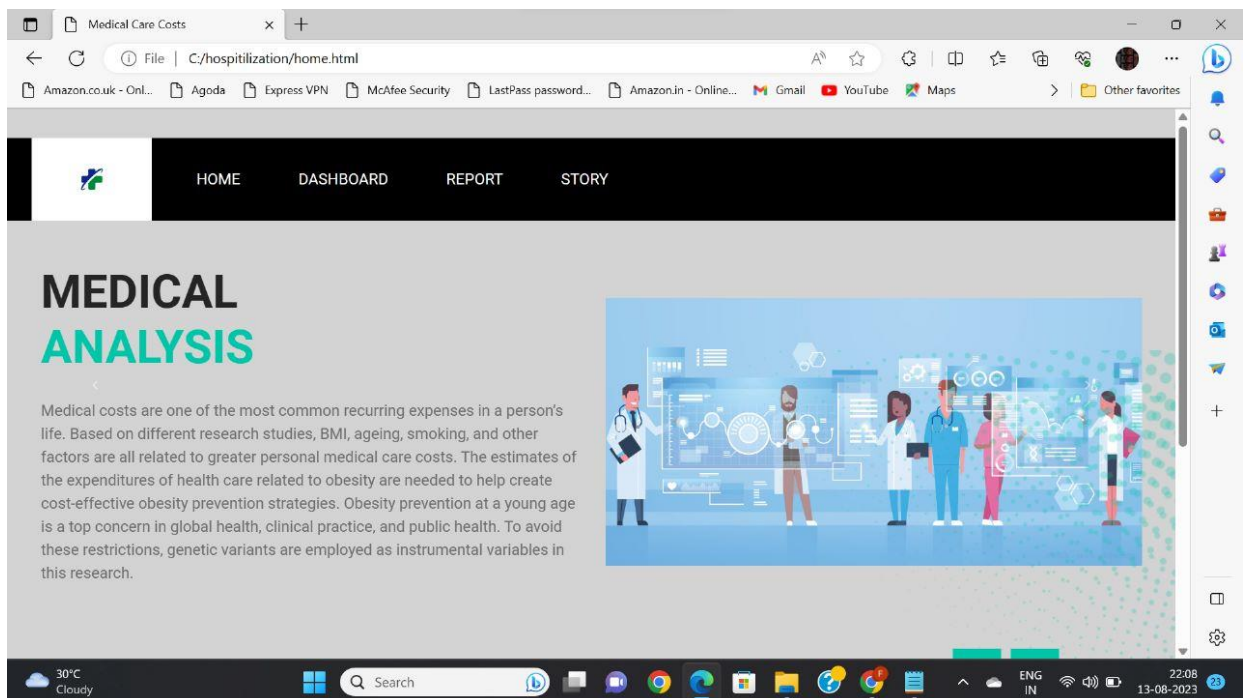
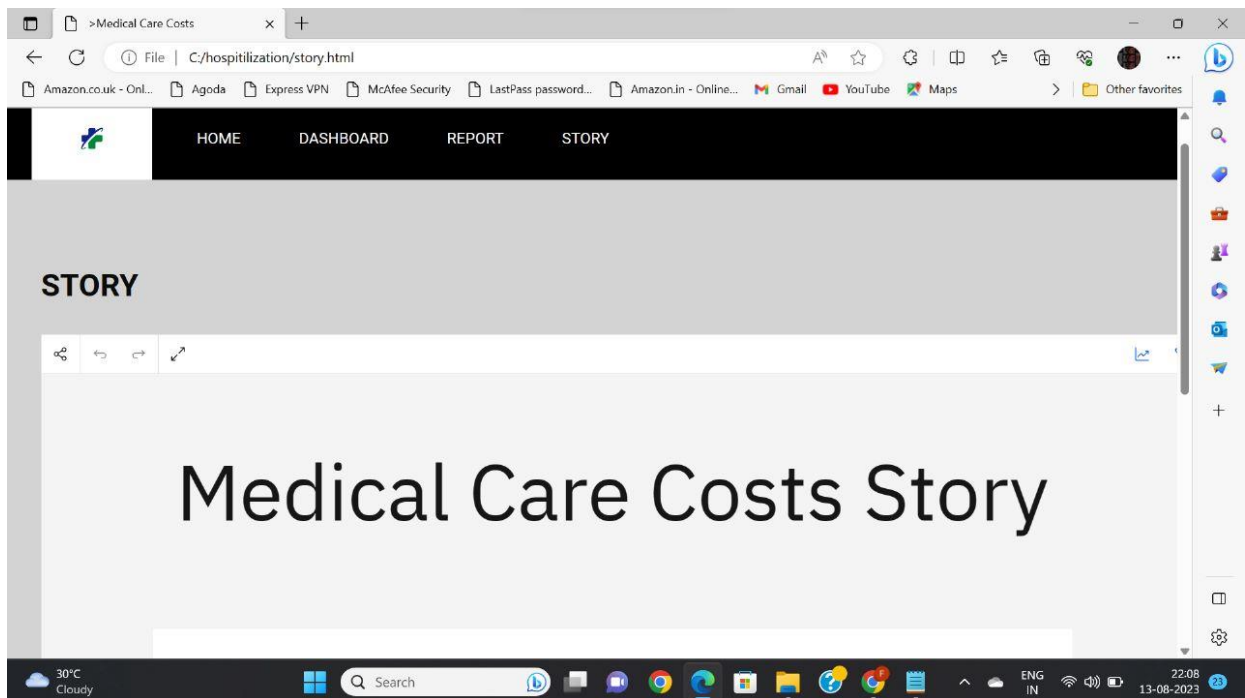
3. The Proposed Method Based on Linear Regression

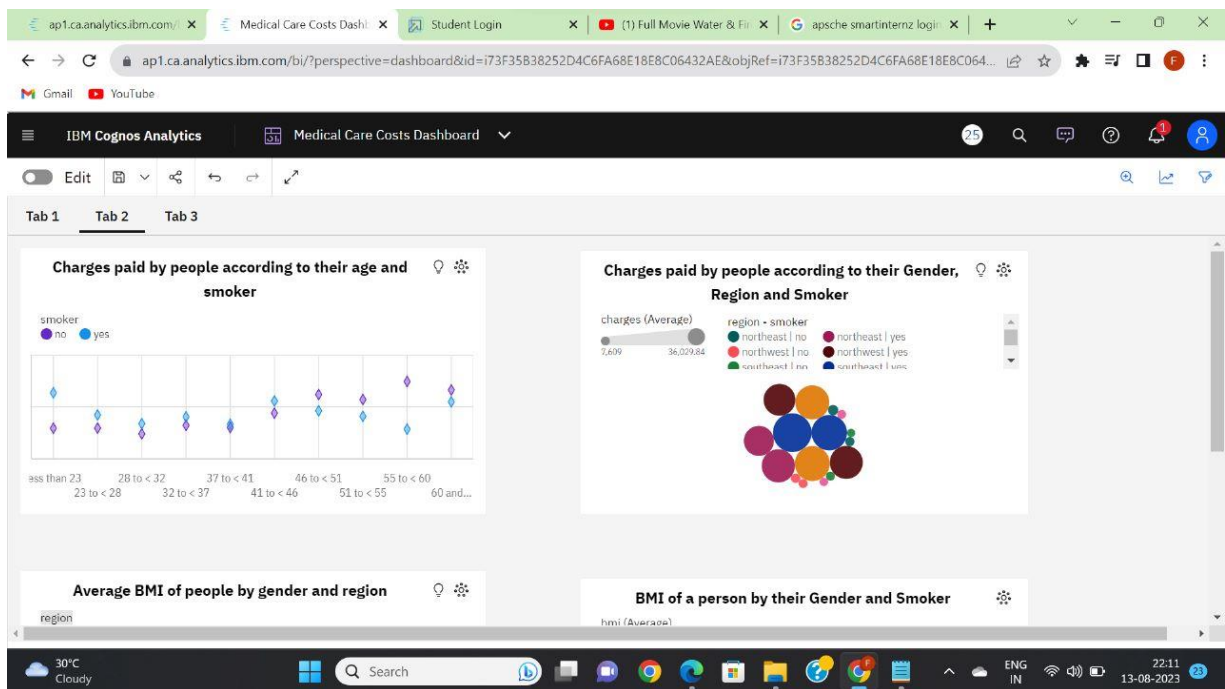
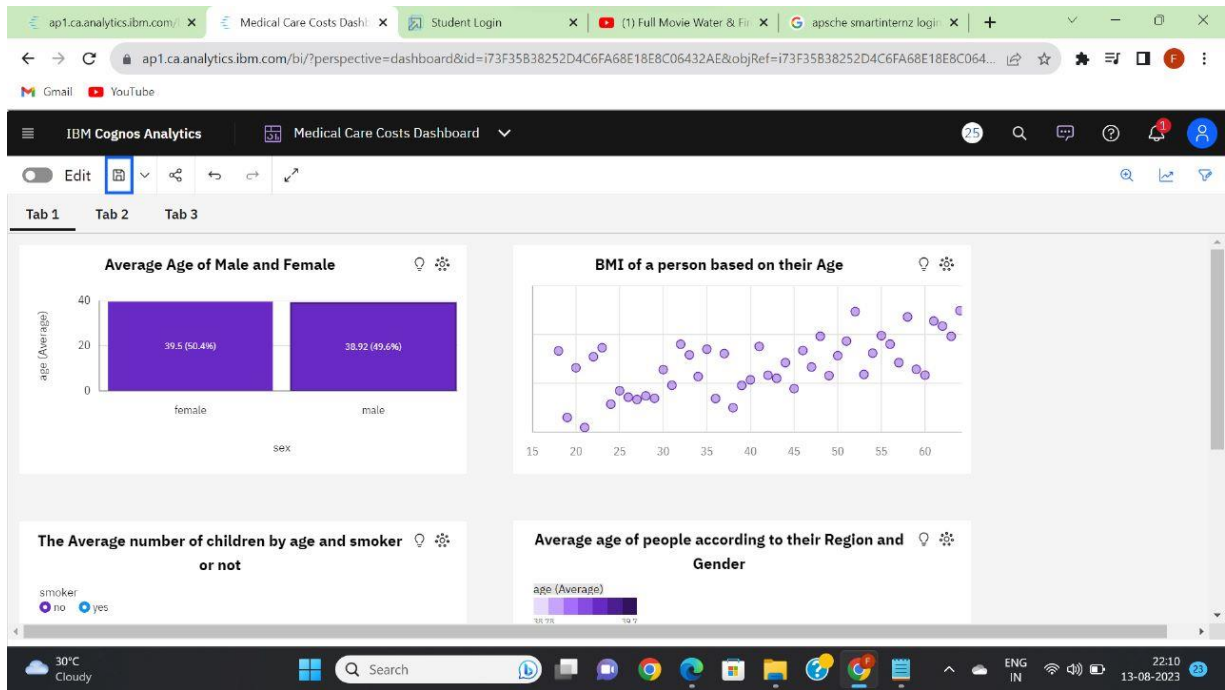
Linear regression is one of the most common supervisory machine learning statistical analysis techniques. It is commonly used to find linear correlations between two or more responses and predictive variables. The technique is divided into two types depending on the number of variables in the model such as simple linear regression and multiple linear regression. A response variable corresponding to a predictive variable is simple linear regression. Whether more than two response variables correspond to predictive variables is known as multiple linear regression as shown in Figure 1. This work used linear regression to study the relationship among total maintenance and other properties in datasets to obtain the properties most affected by the total cost of maintenance. 75% of the data in the dataset were trained, and 25% of the data were tested. Then, Pearson's correlation coefficient (PCC) for each simple linear regression sample was calculated. The PCC is determined and calculated by the following equation to find the parallel variability and strength of a linear regression relationship between two factors:

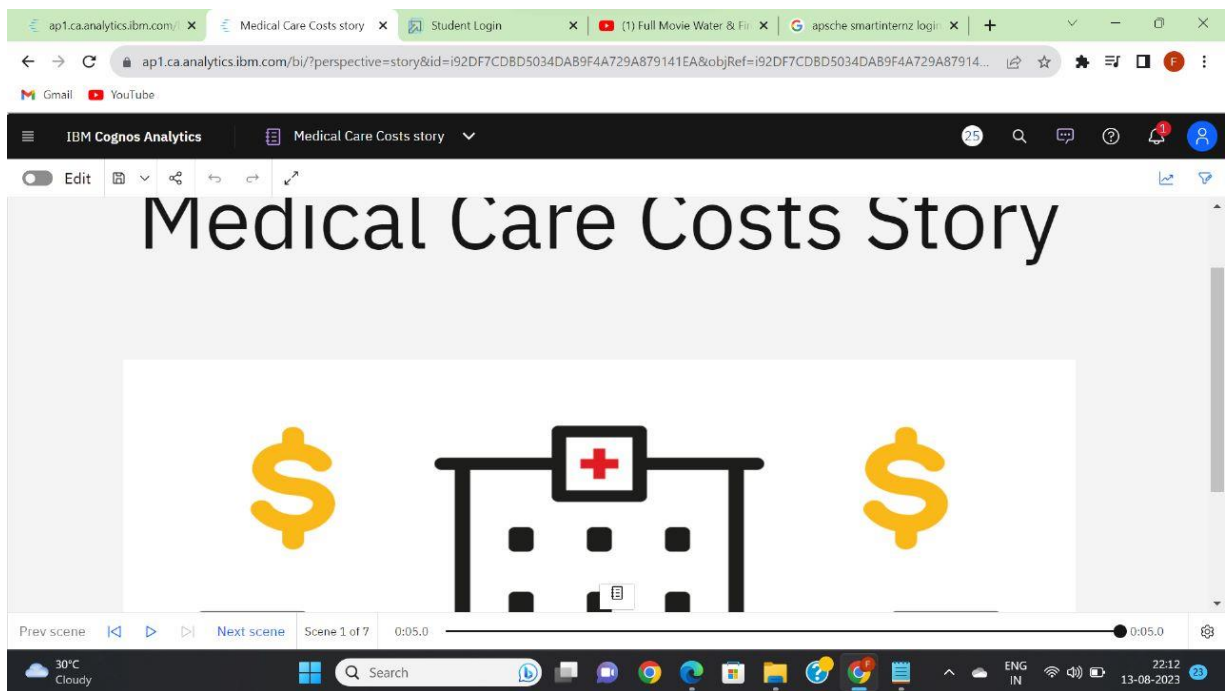
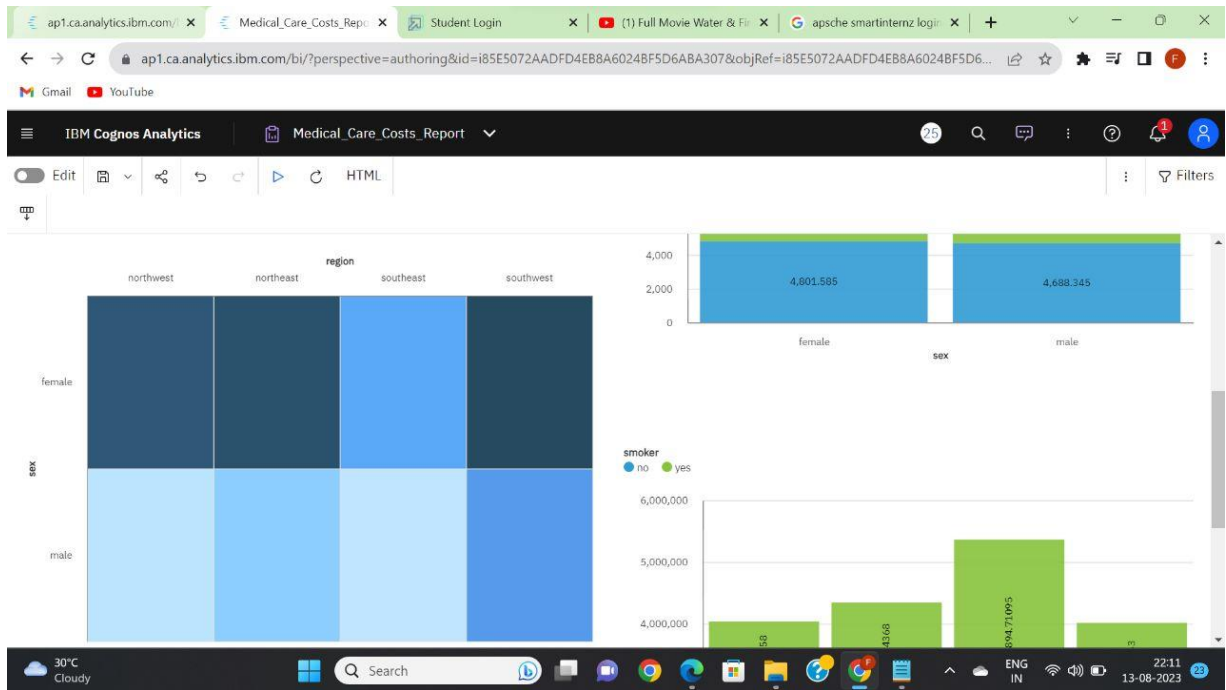


CHAPTER 4 EXPERIMENTAL RESULTS









CHAPTER 5 : APPLICATIONS/ADVANTAGES

ADVANTAGES

1. Easy access to all health care policies
2. Distance dont matter to know consultants
3. Reduced costs including offers
4. Facilitated and secured payment processes
5. Enhanced workflow for healthcare professionals
6. Ensures Assurance and Convenience for patients

DIS ADVANTAGES

1. Technical Issues and Data concerns
2. Not Appropriate during lack of smartphones and technical knowledge
3. Privacy policies may lag due to network issues
4. Showing wrong results and errors
5. Lack of doctor patient interactions

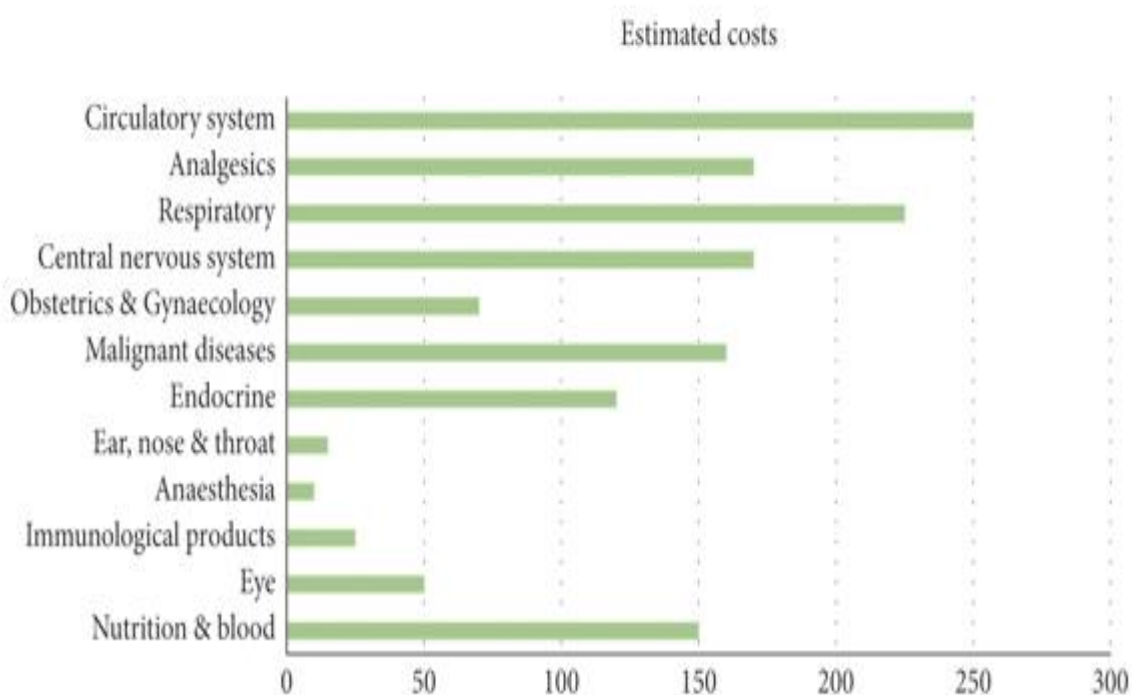
FEATURES:

1. Drug Discovery and Manufacturing
2. Medical Imaging Diagnosis
3. Behavioral Modification
4. Smart Health Records
5. Clinical Trial and Research
6. Crowdsourced Data Collection
7. Better Radiotherapy
8. Outbreak Prediction

CHAPTER 6 : Results and Analysis:

The average annual rates and costs of consultations, tests, and prescription items were estimated by BMI category at the time of recruitment as shown in Figure 4. Percentage differences in rates and average annual costs were calculated for women with a BMI greater than 2 kg/m² and a BMI greater than 20 kg/m², both overall and according to the type of drug use.

All models were evaluated using semi-possible generalized linear models with variations such as record link and Poisson. At the beginning of each year, annual expenses are estimated in subgroups defined by alcohol consumption, socioeconomic status, smoking level, educational qualifications, and strenuous exercise in recruitment. The diversity of the proportional increases in annual costs among the types of each subgroup was estimated using the chi-square test.



Healthcare expenses attributable to obesity and overweight between people on a yearly basis.

CHAPTER 7: CONCLUSIONS

We provided a new linear regression that can easily demonstrate the reasons for producing a certain forecast regarding potential healthcare expenses, which is a useful capacity in the healthcare area. The linear regression algorithm is used to estimate the healthcare costs of the patients such as obesity (BMI) using certain devices such as smartphones and smart devices. For estimation, by the use of linear regression, supervised learning performs more accurately. By providing comprehensive evidence, regression methodology can be effectively used for prognosis in conjunction with the dataset. The domain and time accuracy will determine the prediction model and the estimation of healthcare expenses. The proposed method reduces the risk of overfitting, and also, training time is less. This method is effective in estimating the healthcare costs of patients with an accuracy rate of 97.89%. The extensive tests on a real-time world database have confirmed the efficiency of our method

Inference

IBM Database extracted from smart internz.
Google.