A Project Report on

Estimation and Prediction of Hospitalization and Medical Care Costs

by

Team Leader: GADDAM. INDRA NANDHA(21AT5A0504)
Student 1: PINJARI SHABANA BEGUM(20AT1A05D4)
Student 2: DALAVAI PAMPAPATHI (21AT5A0503)
Student 3: YERRANAGU HARIBHASKAR REDDY (21AT5A0518)

Under the Guidance of

Mrs.M.Jaya Sunitha M.Tech*
Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND 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)

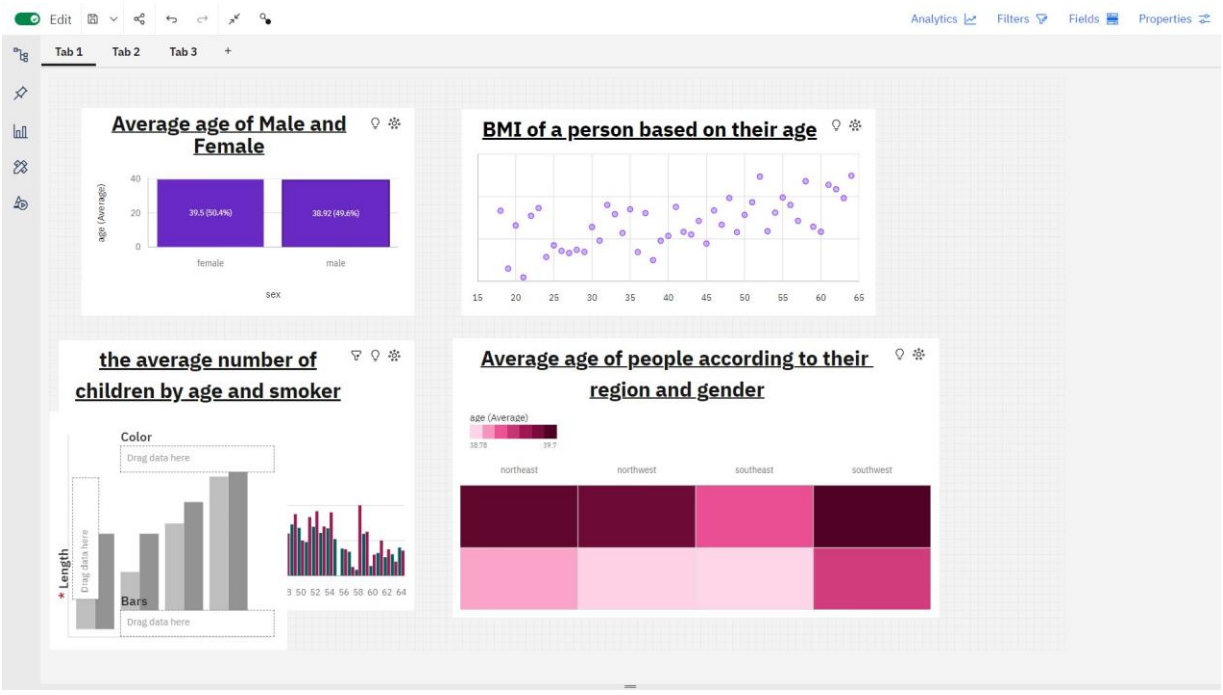
ABSTRACT

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

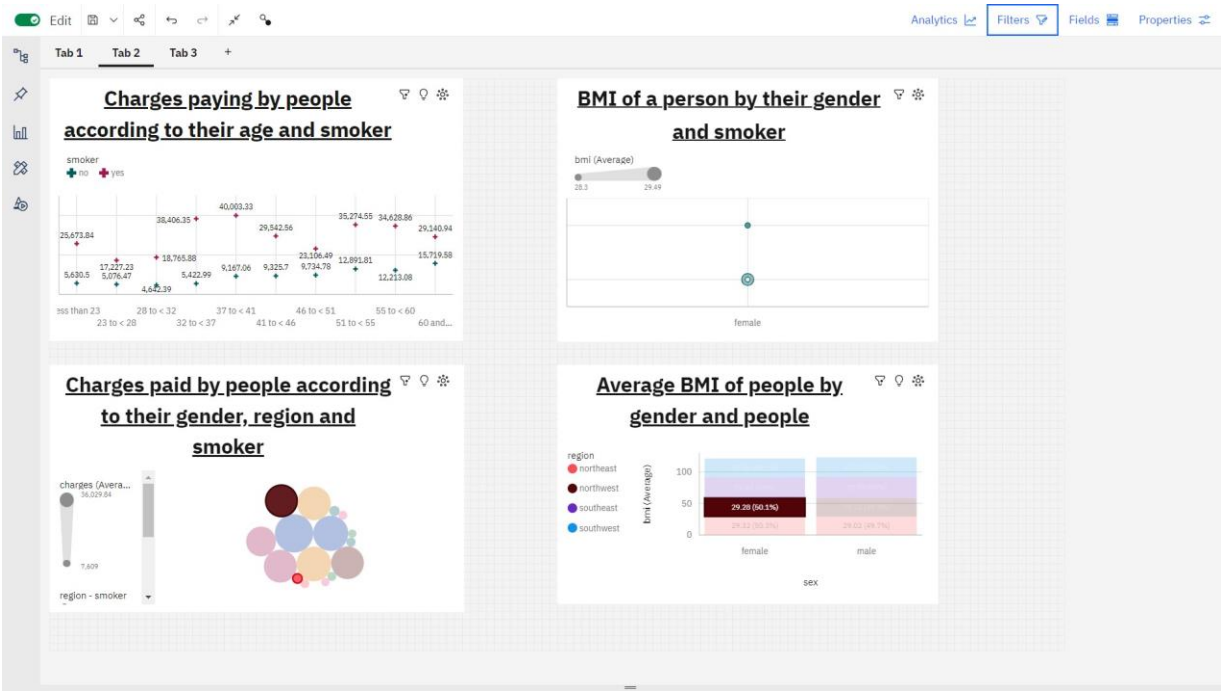
Medical costs are one of the most common recurring expenses in a person's life. Based on different research studies, BMI, ageing, smoking, and other factors are all related to greater personal medical care costs. .e estimates of the expenditures of health care related to obesity are needed to help create 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 machinelearning 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

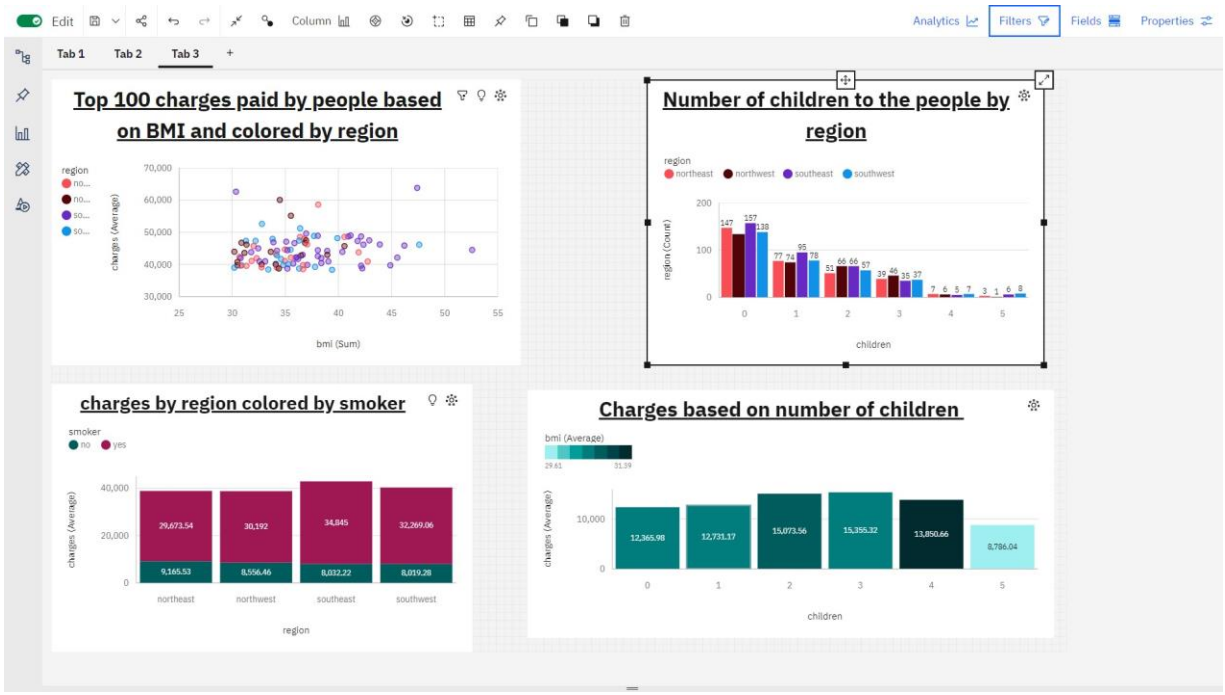
LIST OF FIGURES ANDTABLES



Dashboard 1

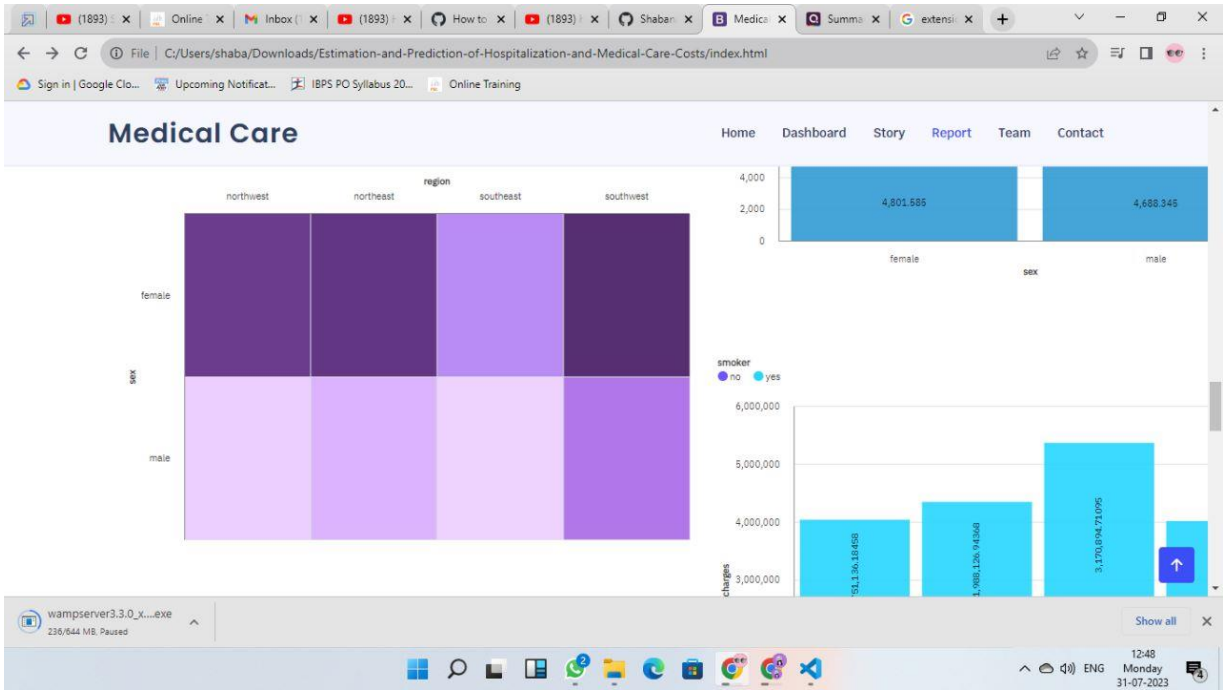


Dashboard 2



Dashboard 3

Story



Report

Web Integration

CHAPTER 1
INTRODUCTION

The incidence of overweight and obesity has increased significantly in most countries in recent decades. Excess weight is associated with an increas

ed 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, conditions, 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

2.0 Data Collection & Extraction From Database

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes and generate insights from the data.

2.1 Collect The Dataset

Understand the data

Data contains all the meta information regarding the columns described in the CSV files. we have provided two CSV file:

- health_events.csv
- noc_regions.csv

Column Description for pepole_events.csv:

The file pepole _events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual pepole competing in an individual Health event (athlete-events). The columns are:

- ID: Unique identifier for each pepole
- Name: Name of the pepole
- Sex: Gender of the pepole (M/F)
- Age: Age of the pepole
- Height: Height of the pepole
- Weight: Weight of the pepole in kilograms
- Team: Name of the country the pepole represents
- NOC: Three-letter code of the country the pepole represents
- conditions: Year and season of the Health conditions (e.g., "2000 Summer")
- Year: Year of the Health conditions
- Season: Season of the Health conditions (Summer/Winter)
- City: Name of the city where the Health conditions were held
- state: state the pepole participated in
- Event: Specific event the pepole participated in
- Medal: Type of medal the pepole won (Gold/Silver/Bronze)

Column Description for noc_regions.csv:

- NHC: Three-letter code of the National Health Committee
- Country: Name of the country represented by the NHC
- Notes: Additional notes about the NOC or country

2.2 Storing Data In DB2 & Perform SQL Operations

In this activity we will see hot to store data in DB2

2.3 Connect DB2 With Cognos

In this activity, we will seen how to connect IBM DB2 and cognos analytics

CHAPTER 3

3.0 Data Preparation

In this milestone, we will see how to prepare the data for building visualizations

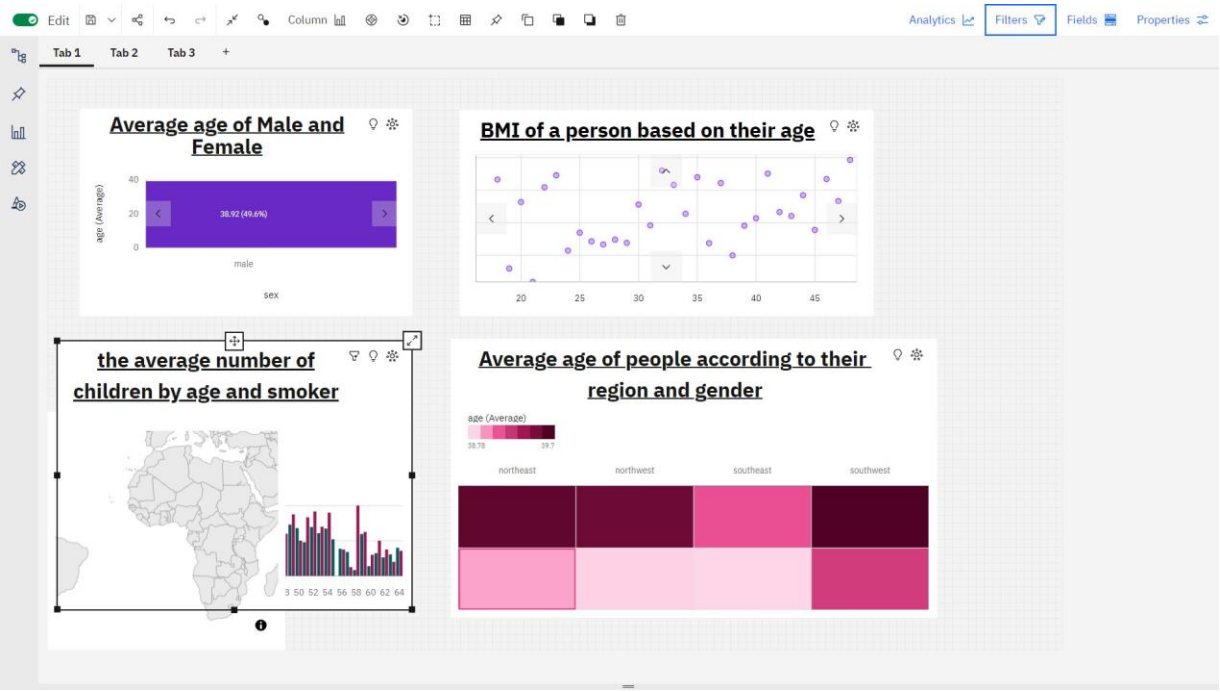
3.1 Prepare The Data For Visualization

Preparing the data for visualization involves cleaning the data to remove irrelevant or missing data, transforming the data into a format that can be easily visualized, exploring the data to identify patterns and trends, filtering the data to focus on specific subsets of data, preparing the data for visualization software, and ensuring the data is accurate and complete. This process helps to make the data easily understandable and ready for creating visualizations to gain insights into the performance and efficiency.

CHAPTER 4

4.0 Data Visualization

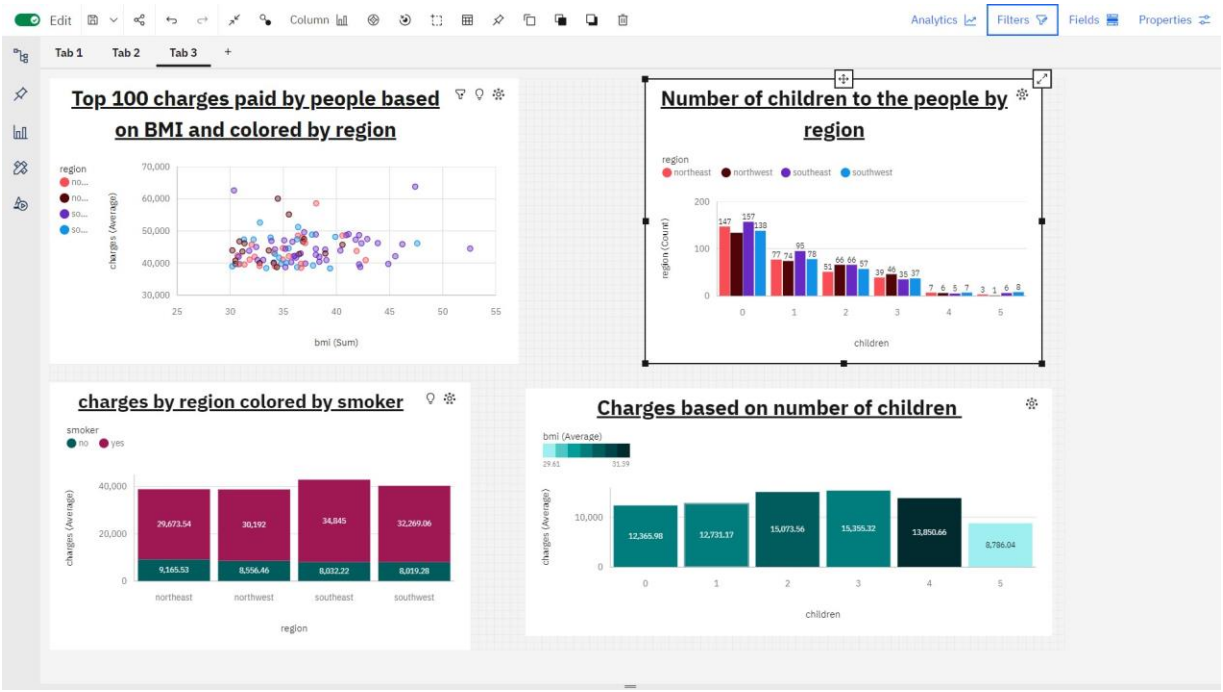
Data visualization is the process of creating graphical representations of data in order to help people understand and explore the information. The goal of data visualization is to make complex data sets more accessible, intuitive, and easier to interpret. By using visual elements such as charts, graphs, and maps, data visualizations can help people quickly identify patterns, trends, and outliers in the data.

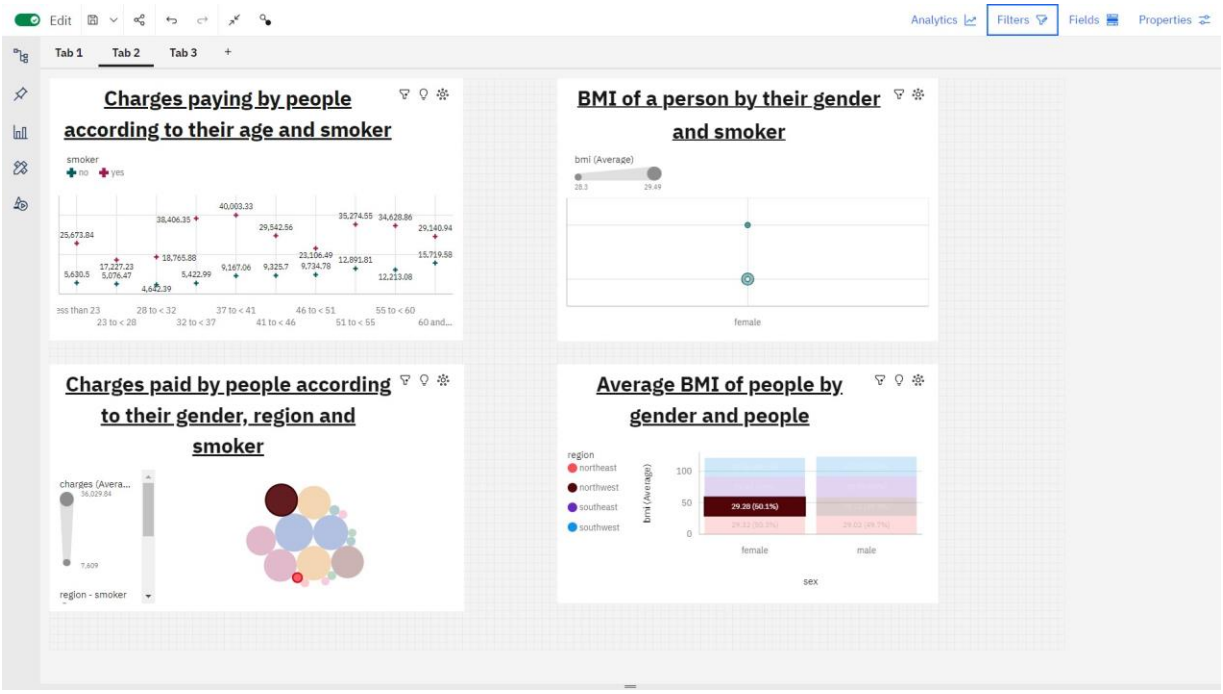


CHAPTER 5

5.0 Dashboard

A dashboard is a graphical user interface (GUI) that displays information and data in an organized, easy-to-read format. Dashboards are often used to provide real-time monitoring and analysis of data, and are typically designed for a specific purpose or use case. Dashboards can be used in a variety of settings, such as business, finance, manufacturing, healthcare, and many other industries. They can be used to track key performance indicators (KPIs), monitor performance metrics, and display data in the form of charts, graphs, and tables.


Edit
Column
Analytics
Filters
Fields
Properties
Tab 1
Tab 2
Tab 3
Top 100 charges paid by people based on BMI and colored by region
region
charges (Average)
bmi (Sum)
Number of children to the people by region
region
northeast
northwest
southeast
southwest
region (Count)
children
charges by region colored by smoker
smoker
no
yes
charges (Average)
29,673.54
30,192
34,845
32,269.06
9,165.53
8,556.46
8,032.22
8,019.28
northeast
northwest
southeast
southwest
region
Charges based on number of children
bmi (Average)
charges (Average)
12,365.98
12,731.17
15,073.56
15,355.32
13,850.66
8,786.04
children


Edit
Analytics
Filters
Fields
Properties
Tab 1
Tab 2
Tab 3
Charges paying by people according to their age and smoker
smoker
no
yes
BMI of a person by their gender and smoker
bmi (Average)
female
Charges paid by people according to their gender, region and smoker
charges (Avera...
region - smoker
Average BMI of people by gender and people
region
northeast
northwest
southeast
southwest
bmi (Average)
29.28 (50.1%)
female
male
sex

CHAPTER 6

6.0 Story

A data story is a way of presenting data and analysis in a narrative format, with the goal of making the information more engaging and easier to understand. A data story typically includes a clear introduction that sets the stage and explains the context for the data, a body that presents the data and analysis in a logical and systematic way, and a conclusion that summarizes the key findings and highlights their implications. Data stories can be told using a variety of mediums, such as reports, presentations, interactive visualizations, and videos.

6.1 No Of Scenes Of Story

The number of scenes in a storyboard for Data-Driven insights on Health conditions Participation and Performance will depend on the complexity of the analysis and the specific insights that are trying to be conveyed. A storyboard is a visual representation of the data analysis process and it breaks down the analysis into a series of steps or scenes.

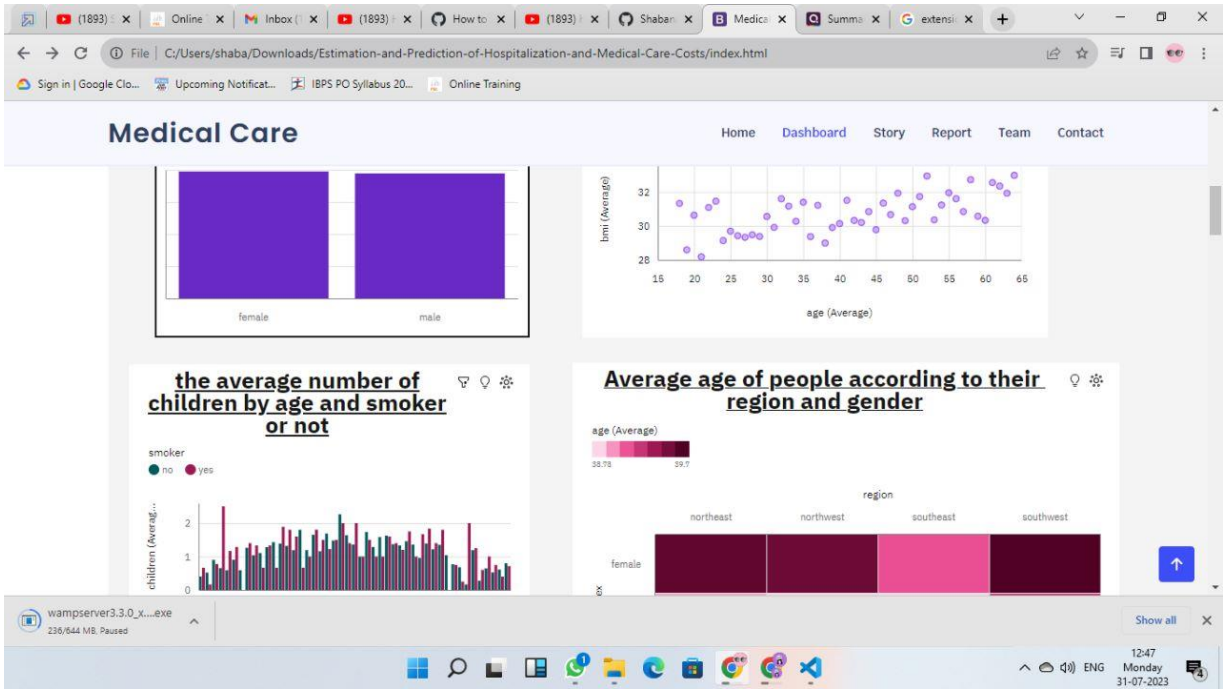
CHAPTER 7

7.0 Report

A report in data analytics typically involves analyzing and interpreting data to draw insights and conclusions that can inform business decisions or address research questions. The report usually includes a summary of the data analysis process, including the methods and tools used, as well as the findings and recommendations based on the analysis. The report should begin with an executive summary, which provides a brief overview of the main findings and recommendations. The introduction should provide background information on the problem or research question being addressed and the data sources used.

7.1 No.Of Visualization With Detail Information

When creating a report in cognos, it is often helpful to include visualizations to help communicate the findings of the analysis.



CHAPTER 8

8.0 Performance Testing

8.1 Amount Of Data Rendered To DB2

The amount of data that is rendered to a database depends on the size of the dataset and the capacity of the database to store and retrieve data

CHAPTER 9

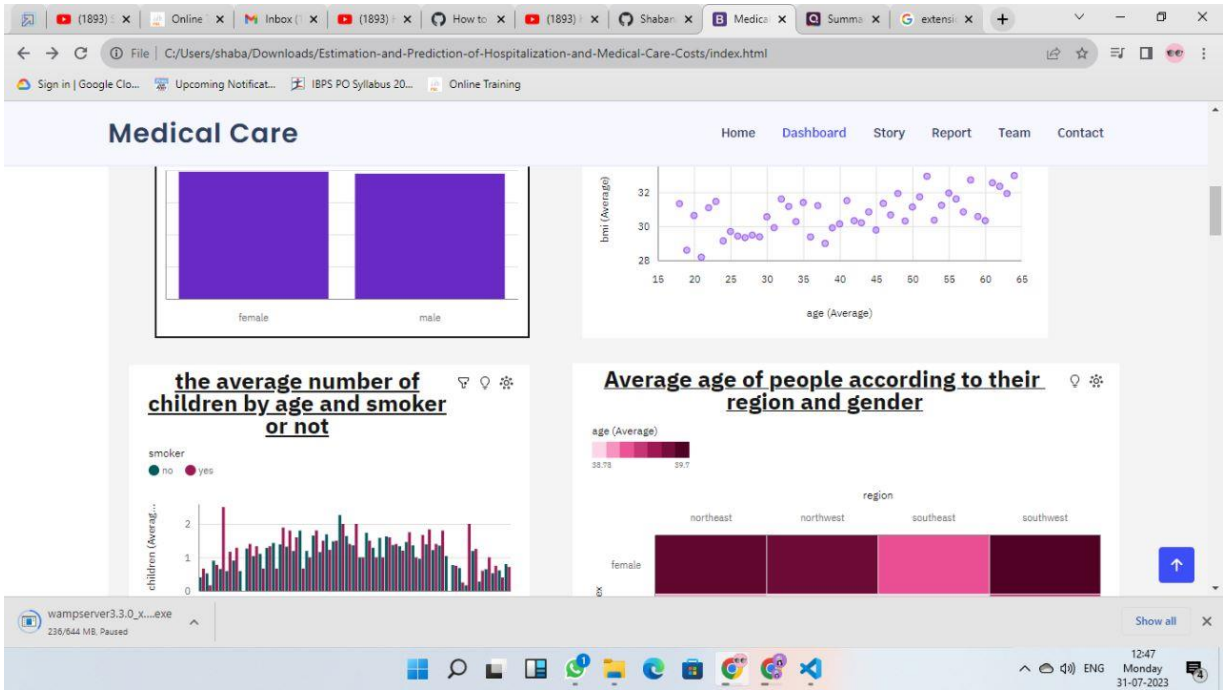
9.0 Web Integration

- Publishing helps us to track and monitor key performance metrics, to communicate results and progress. help a publisher stay informed, make better decisions, and communicate their performance to others.

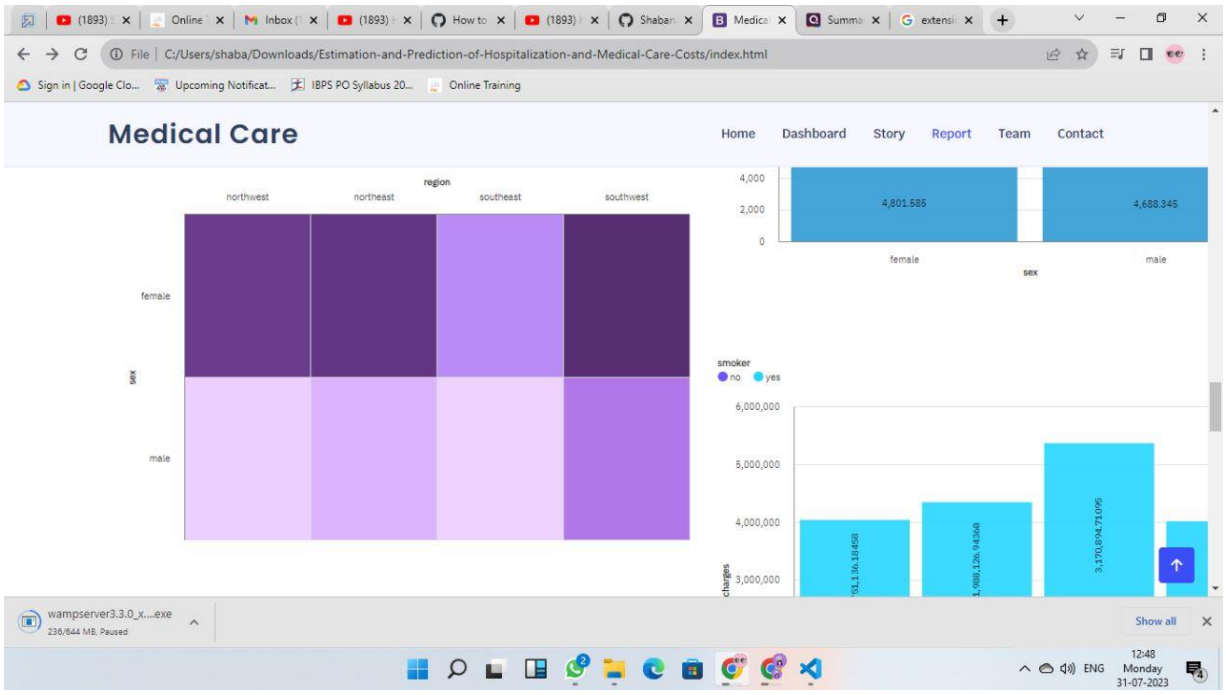
9.1 Publishing dashboard, report & story.

- Step 1: Go to Dashboard, report & /story, click on share button on the top.
- Dashboard

HOME PAGE

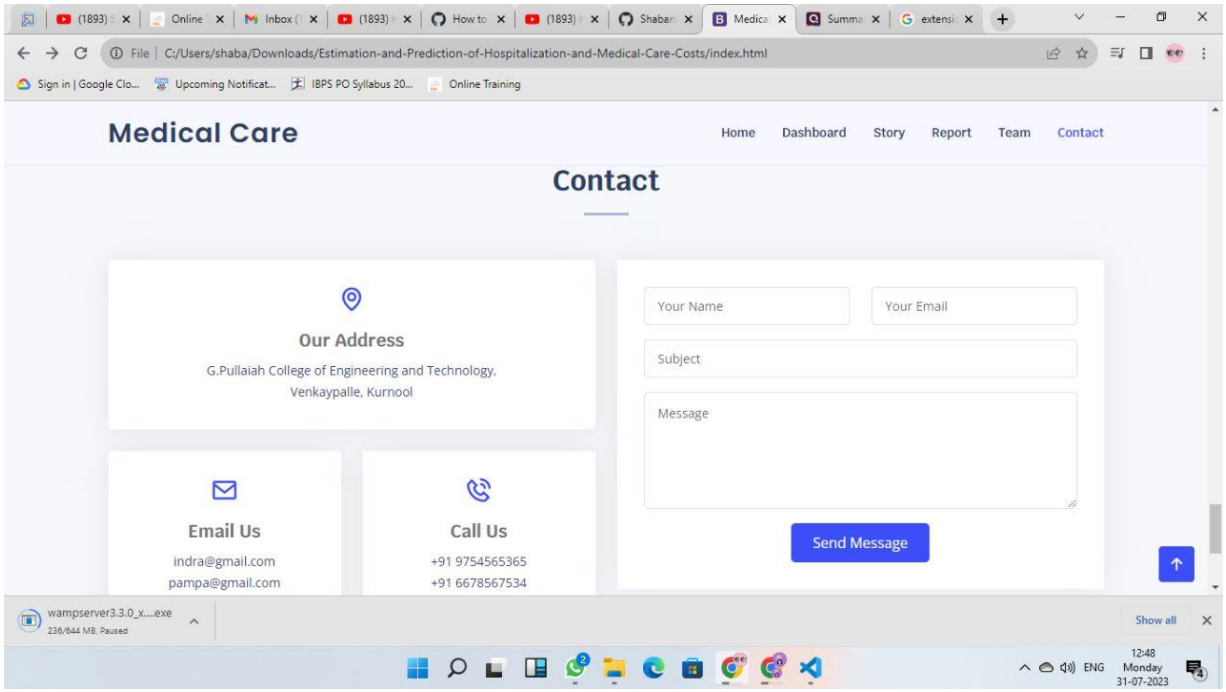


DASHBOARD

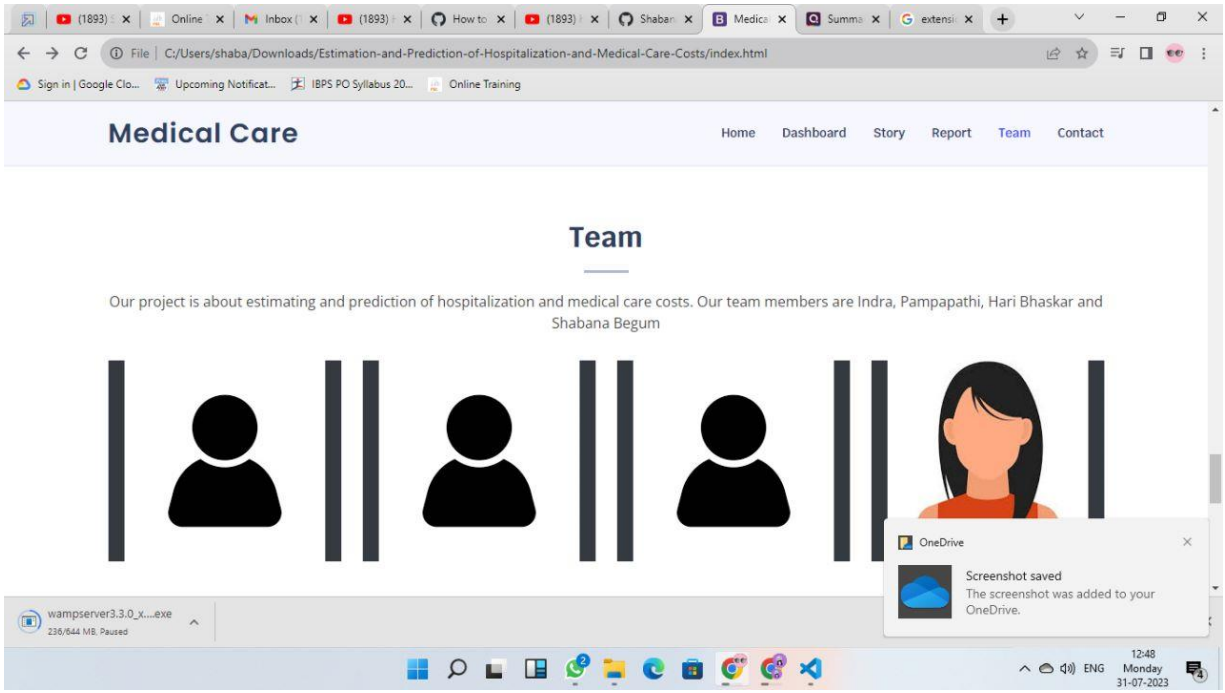


REPORT

STORY



Contact



Team