**Estimation and Prediction of Hospitalization and**

**Medical Care Costs**

**TEAM ID : NM2023TMID11307**

|  |  |
| --- | --- |
| **TEAM LEAD** | SUDARSHANA V |
| **TEAM MEMBER - 1** | YESSWINE R |
| **TEAM MEMBER - 2** | DEEPIKA S |
| **TEAM MEMBER - 3** | KAUSIKA N G |

1. **INTRODUCTION**

The price of hospitalisation and medical care has significantly increased recently, which has had an impact on the healthcare sector. The burden of this growth has been felt by everyone, including patients, insurance companies, and healthcare systems. Data analytics has become a potent tool for measuring and predicting hospitalisation and medical care expenditures in order to tackle this problem.

To find significant insights and trends, data analytics entails the extraction, interpretation, and analysis of huge, complex datasets. Data analytics helps healthcare providers, insurers, and politicians make well-informed decisions, allocate resources efficiently, and create cost-containment plans that work..

The price of hospitalisation and medical care has significantly increased recently, which has had an impact on the healthcare sector. The burden of this growth has been felt by everyone, including patients, insurance companies, and healthcare systems. Data analytics has become a potent tool for measuring and predicting hospitalisation and medical care expenditures in order to tackle this problem.

To find significant insights and trends, data analytics entails the extraction, interpretation, and analysis of huge, complex datasets. Data analytics helps healthcare providers, insurers, and politicians make well-informed decisions, allocate resources efficiently, and create cost-containment plans that work , and ultimately cut expenditures by better managing or avoiding chronic illnesses.

Data analytics requires the extraction, interpretation, and analysis of enormous, complicated datasets in order to identify important insights and patterns. Data analysis is useful.In conclusion, data analytics have enormous promise for enhancing healthcare cost management through the estimation and forecast of hospitalisation and medical care expenses. Healthcare stakeholders can identify high-risk individuals and acquire insightful knowledge on cost drivers by utilising the power of data. This information can help with decision-making, resource allocation, and the creation of cost-containment plans, which will ultimately result in more effective and long-lasting healthcare systems..

**1.1 Project Overview**

This initiative aims to evaluate and forecast the expenses of hospitalisation and medical care by utilising data analytics techniques. The project intends to uncover cost causes, create prediction models, and offer insights for efficient cost management in the healthcare industry by analysing huge and complicated healthcare information.

**Methodology:**

**Data Collection:** Gather relevant datasets containing information on patient demographics, diagnoses, procedures, length of hospital stay, medication usage, and associated costs. This data may be obtained from healthcare providers, insurance companies, and public health agencies.

**Data Preprocessing:** Cleanse and preprocess the collected data to ensure accuracy and consistency. Handle missing values, eliminate outliers, and standardize data formats for further analysis.

**Exploratory Data Analysis:** Conduct exploratory data analysis to gain insights into the underlying patterns, trends, and relationships within the data. Identify potential cost drivers and variables that significantly impact hospitalization and medical care costs.

**Feature Engineering:** Create new features or transform existing variables to enhance the predictive power of the models. This may involve feature scaling, dimensionality reduction, or creating derived variables that capture meaningful information.

**Model Development:** To create predictive models, use the proper data analytics techniques, such as regression analysis, machine learning approaches, or time series forecasting. Utilising historical data with known costs, train the models, and then assess their performance using the right criteria.

Estimate the costs of particular medical procedures, treatments, or illnesses using the provided models. To provide precise cost estimates, consider patient characteristics and pertinent variables.

**Cost Prediction:** Use the predictive algorithms to project future cases' hospitalisation and healthcare expenditures. These forecasts can help with financial planning, resource allocation, and budgeting for healthcare providers, insurance providers, and policymakers.

**1.2 Purpose**

The estimation and forecasting of hospitalisation and medical care expenses in data analytics has a variety of uses and contributes to a number of significant goals in the healthcare sector. The main goals are as follows:

**Cost Control:** The main goal is to make it possible for the healthcare industry to manage costs effectively. Healthcare providers, insurance providers, and policymakers may create strategies to reduce costs, optimise resource allocation, and make wise financial decisions by precisely assessing and projecting hospitalisation and medical care costs.

**Financial Planning:** Cost estimation and forecasting aid people, insurance companies, and healthcare systems in making financial plans. Patients can more clearly comprehend and prepare for the financial costs associated with medical procedures or treatments, and insurance companies can set reasonable premiums and coverage thresholds. Healthcare systems are capable of effectively allocating funds and resources.

Data analytics can be used to find the cost-drivers and risk factors connected to hospitalisation and medical care. Predictive algorithms can identify people who are more likely to require expensive medical treatment by examining historical data and patient characteristics. In order to lower hospitalisation rates and related costs, healthcare practitioners can use this information to implement preventative measures, proactive interventions, and disease management strategies. Cost estimation and forecasting can also be used to assess how well healthcare systems, insurance providers, and providers themselves are performing. Stakeholders can evaluate the efficacy of their cost management methods, pinpoint areas for improvement, and track the financial impact of interventions and initiatives by comparing actual costs with expected costs.

1. **IDEATION & PROPOSED SOLUTION**

**IDEATION:**

For healthcare providers, insurers, and policymakers, the estimation and forecasting of hospitalisation and medical care costs using data analytics can be a useful tool. It is feasible to create models that can precisely estimate and predict the expenses related to hospitalisation and medical care by utilising historical data, patient information, and other pertinent variables. Decision-making, resource allocation, and financial planning are just a few of the uses for this data.

**PROPOSED SOLUTION**

**Data collection:** Compile thorough and pertinent information from a range of sources, such as electronic health records, insurance claims, demographic data, and records of medical operations, diagnosis codes, and billing transactions. Make sure the information is anonymised and complies with privacy laws.

**Data Preprocessing:** To manage missing values, outliers, and inconsistencies, clean up and preprocess the obtained data. To prepare the data for analysis, perform data normalisation, feature scaling, and categorical variable encoding.

Select the features that will have the most bearing on the expenditures associated with hospitalisation and medical care. Use methods like feature importance, correlation analysis, and domain expertise to choose the best set of features for modelling.

**Model Development:** Explore various machine learning algorithms suitable for regression tasks, such as linear regression, decision trees, random forests, or gradient boosting. Train and evaluate different models using the preprocessed data, considering appropriate evaluation metrics such as mean squared error (MSE) or mean absolute error (MAE).

**Model Validation and Optimization:** Split the data into training and testing sets to evaluate the performance of the developed models. Apply cross-validation techniques to further assess model stability and generalizability. Optimize hyperparameters using techniques like grid search or Bayesian optimization to improve model performance.

**Interpretability and Explainability:** Ensure that the developed models are interpretable and provide insights into the factors influencing the estimated costs. Techniques such as feature importance analysis, partial dependence plots, and SHAP (SHapley Additive exPlanations) values can aid in understanding the model's decision-making process.

**Deployment & Integration**: Once a suitable model has been created, either integrate it into current healthcare systems or deploy it as an application. Users should have a simple interface to enter patient data and get price estimates. Maintain a constant eye on the model and update it with fresh information to increase accuracy and relevance over time.

**Continuous Evaluation and Improvement:** Evaluate the model's performance on a regular basis and compare its forecasts to actual expenses. To discover areas for development and adjust the model appropriately, collect input from users and domain experts.

**Collaboration and Knowledge Sharing:** Encourage communication and sharing of insights and outcomes from the data analytics solution among data scientists, healthcare practitioners, and policymakers. Better decisions, modified policies, and better use of healthcare resources may result from this.

**2.1 Problem Statement Definition**

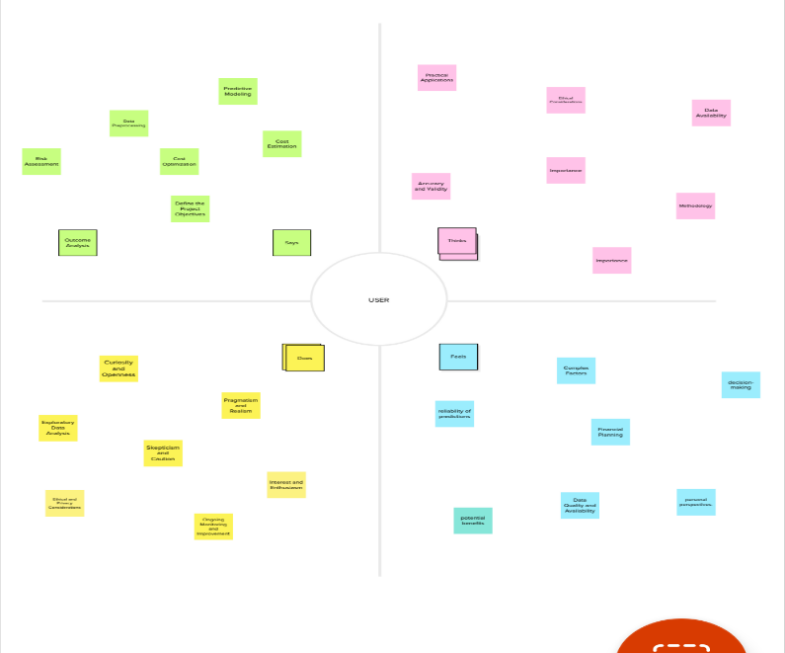
The issue at hand is the estimation and forecasting utilising data analytics of hospitalisation and medical care expenses. The goal is to provide a data-driven system that reliably predicts the costs of hospitalisation and medical care so that healthcare providers, insurers, and policymakers can decide what to do, how to best use their resources, and how much money they should budget for it.

* 1. **Empathy Map Canvas:**

An empathy map is a simple, easy-to-digest visual that captures knowledge about a user’s behaviours and attitudes.

It is a useful tool to helps teams better understand their users.

Creating an effective solution requires understanding the true problem and the person who is experiencing it. The exercise of creating the map helps participants consider things from the user’s perspective along with his or her goals and challenges.



**2.I Ideation & Brainstorming**

Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving. Prioritizing volume over value, out-of-the-box ideas are welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich amount of creative solutions.

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

**Step-1:**

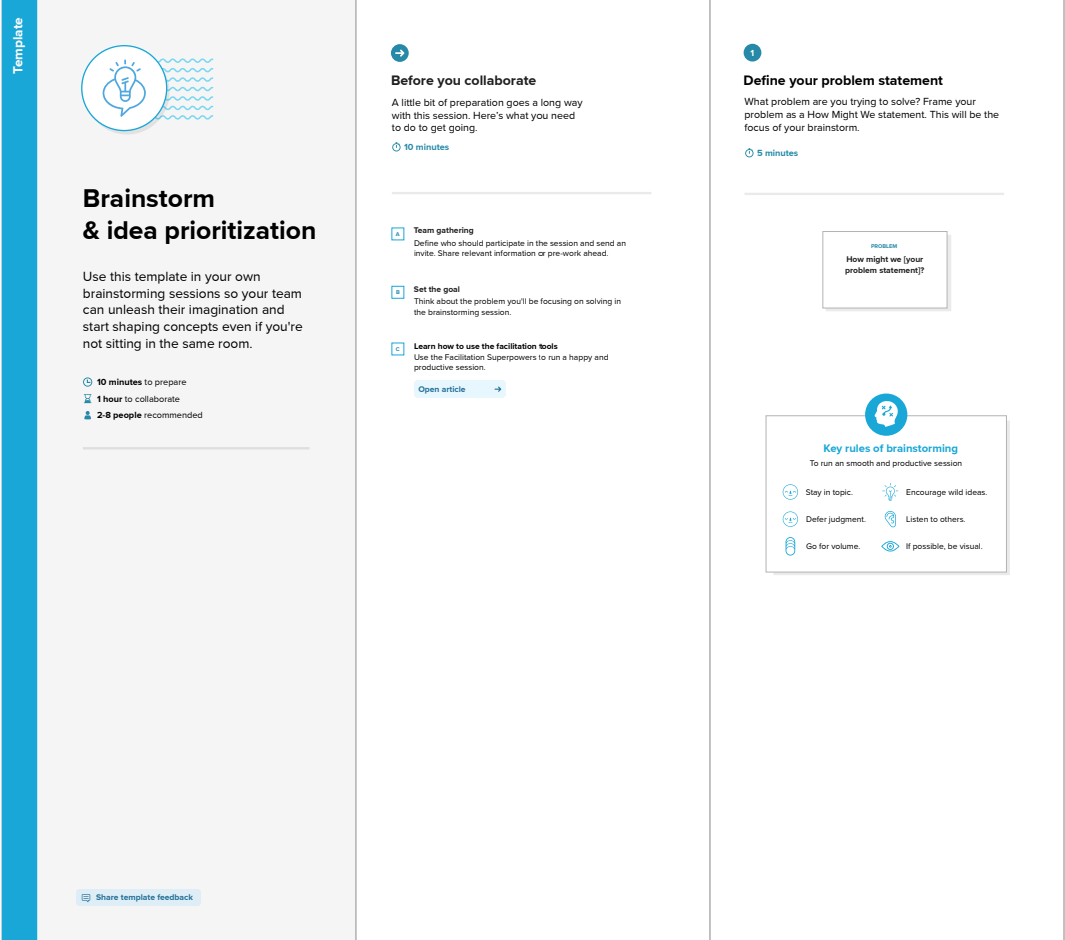


Fig 2: Team Gathering, Collaboration and Select the Problem Statement

**Step-2:**

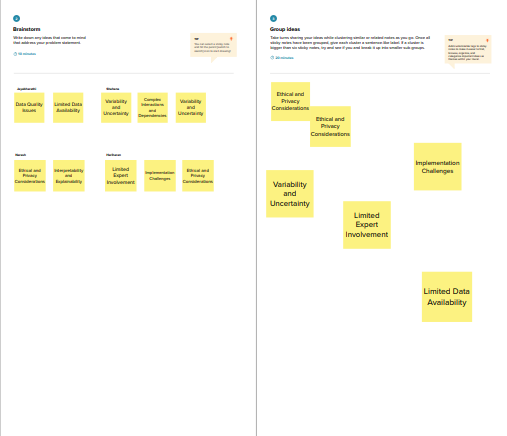
****

Fig 3: Brainstorm, Idea Listing and Grouping

**Step-3:**

****

Fig 4: Idea Prioritization

* 1. **Proposed Solution**

To address the challenges and achieve the project's goals, the following solution is proposed:

* Lack of Transparency
* Inaccurate cost Estimation
* Omplex Billing Structure
* Lack of Persnolized Predection
* Limitted Access to data and tools

By following these steps, you can develop a data-driven solution for estimating and predicting hospitalization and medical care costs, enabling healthcare providers, insurers, and policymakers to make informed decisions and allocate resources more efficiently.

**3.REQUIREMENT ANALYSIS**

**3.1Functional requirement**

The functional requirements for the Estimation and Prediction of Hospitalization and Medical Care Costs data analytics project may include:

* User Registration and Authentication
* Data Input and Management
* Cost Estimation
* Predictive Analytics
* Insurance Integration
* Reporting and Visualization
* Integration with Electronic Health Records (EHR)
* Security and Privacy
* Scalability and Performance
* User Roles and Permissions

It's important to note that the specific functional requirements may vary based on the context and objectives of the student performance analysis system you are envisioning.

**3.2Non-Functional requirements**

Along with the functional requirements, here are some non-functional requirements that can be considered for a student performance analysis system aimed at unleashing the potential of our youth:

* Usability
* Security
* Reliability
* Performance
* Accuracy

These non-functional requirements contribute to the overall effectiveness, reliability, and user satisfaction with the student performance analysis system.

**4.PROJECT DESIGN**

**4.1Data Flow Diagrams**

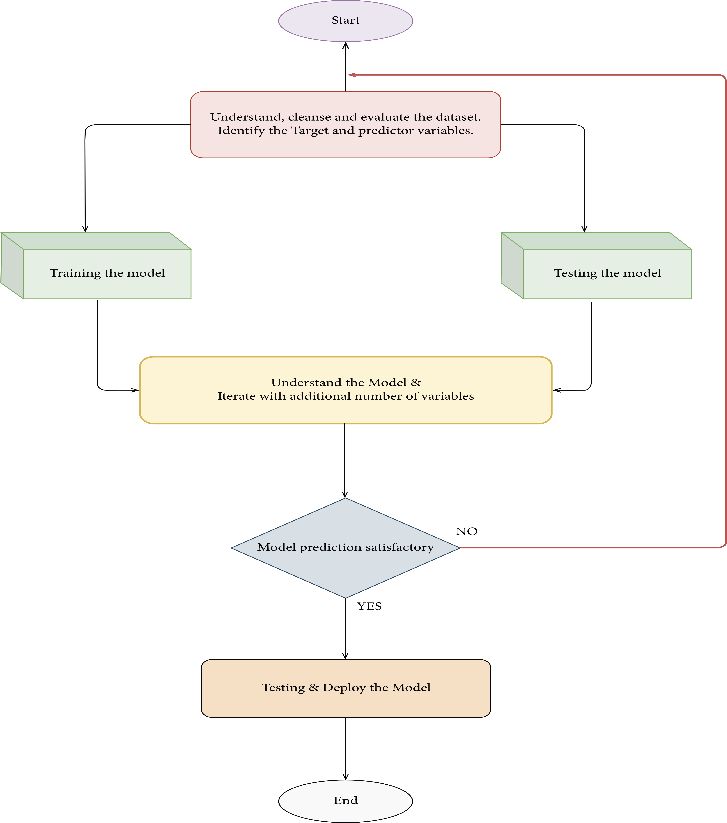
****

Fig 5: Flowchart

**4.2.Solution & Technical Architecture**

**Data Collection and Storage:**

* Gather data from various sources such as electronic health records, insurance claims data, and healthcare databases.
* Store the collected data in a secure and scalable data storage solution, such as a data warehouse or a cloud-based storage system**.**

**Data Preprocessing and Integration:**

* Cleanse and preprocess the data to handle missing values, outliers, and inconsistencies.
* Integrate and consolidate the data from different sources into a unified dataset.
* Perform data normalization or standardization to bring the variables to a common scale and format.

**Data Exploration and Feature Engineering:**

* Conduct exploratory data analysis to gain insights into the data and identify relevant patterns or trends.
* Perform feature engineering to extract meaningful features from the data that can contribute to the prediction of hospitalization and medical care costs.

**Model Development and Training:**

* Select appropriate machine learning or statistical modeling techniques, such as linear regression, decision trees, random forests, gradient boosting, or neural networks.
* Split the dataset into training and testing sets.
* Train the selected model on the training data using algorithms and libraries suitable for the chosen modeling technique.

**Model Evaluation and Optimization:**

* Evaluate the trained model's performance using appropriate evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), or mean absolute error (MAE).
* Perform model optimization by adjusting hyperparameters through techniques like grid search or randomized search to improve the model's predictive accuracy.

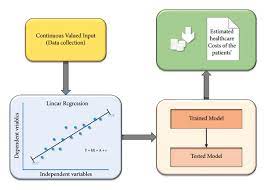
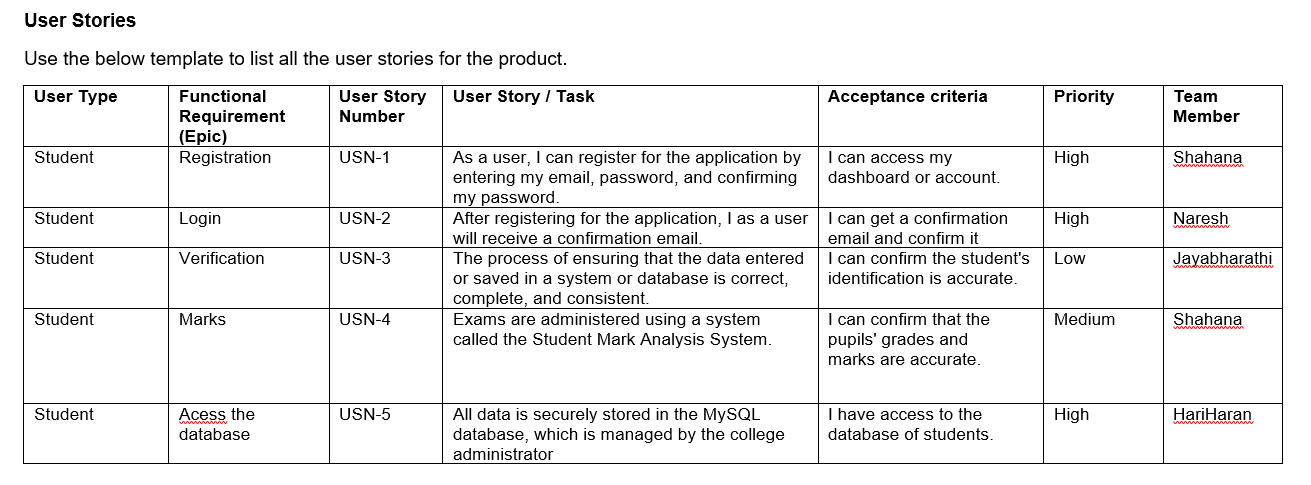


Fig 5: Architecture

**4.3** **User Stories**



5.**CODING & SOLUTIONING**

* 1. **Feature 1**

import pandas as pd

import matplotlib.pyplot as plt

data=pd.read\_csv("/kaggle/input/insurance/insurance.csv")

data.columns

data.head(30)

data.isna().sum()

data['sex'].value\_counts()

data['children'].value\_counts()

data['region'].value\_counts()

plt.plot(data["bmi"].head(10),data["charges"].head(10),marker='o')

#plt.scatter(data['charges'],data['smoker'])

plt.show()

x=[]

y=[]

for index,rows in data.iterrows():

if(rows["sex"]=="male"):

x.append(rows["charges"])

y.append(0)

else:

y.append(rows["charges"])

x.append(0)

print(sum(x)/len(x),sum(x))

print(sum(y)/len(y),sum(y))

plt.scatter(data["smoker"],data["charges"])

plt.scatter(data["children"],data["charges"])

plt.scatter(data['sex'],data['charges'])

data["sex"]=data["sex"].map({'male':0,'female':1})

data["smoker"]=data["smoker"].map({'no':0,'yes':1})

dummy=pd.get\_dummies(data["region"],prefix='region')

data=pd.concat([data,dummy],axis=1)

data.drop("region",inplace=True,axis=1)

x=data.drop("charges",axis=1)

y=data["charges"]

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import GradientBoostingRegressor

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x,y, test\_size=0.2, random\_state=2)

gb\_reg = GradientBoostingRegressor(n\_estimators=100, learning\_rate=0.3, max\_depth=2, random\_state=0)

gb\_reg.fit(X\_train, y\_train)

y\_pred = gb\_reg.predict(X\_test)

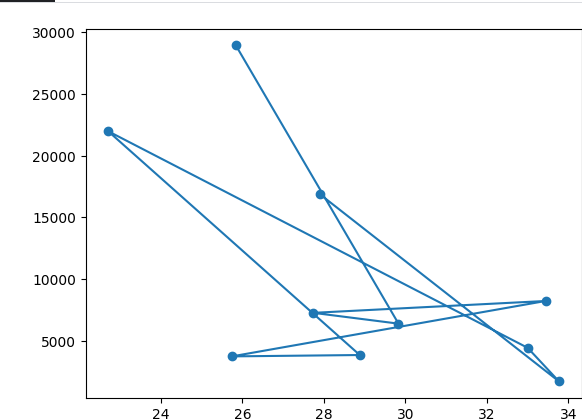
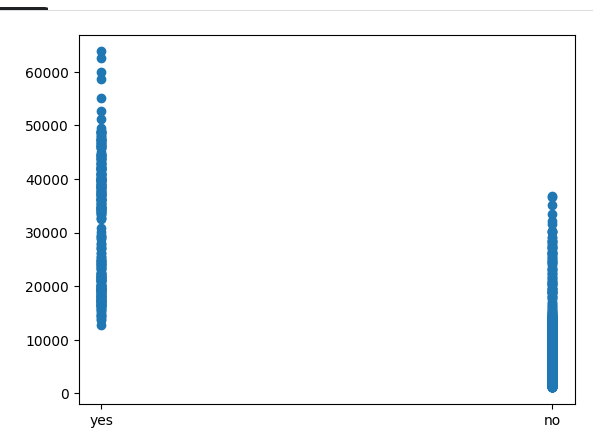
from sklearn.metrics import r2\_score, mean\_squared\_error

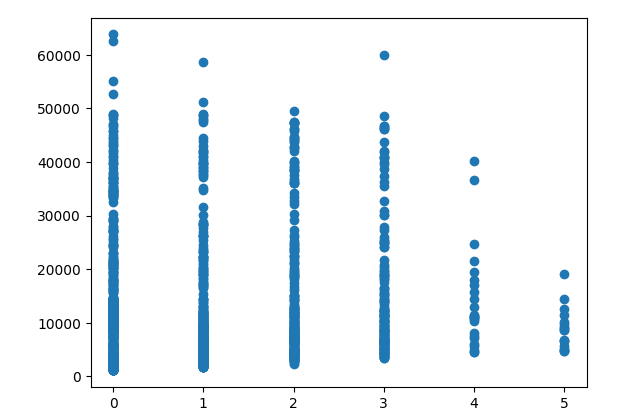
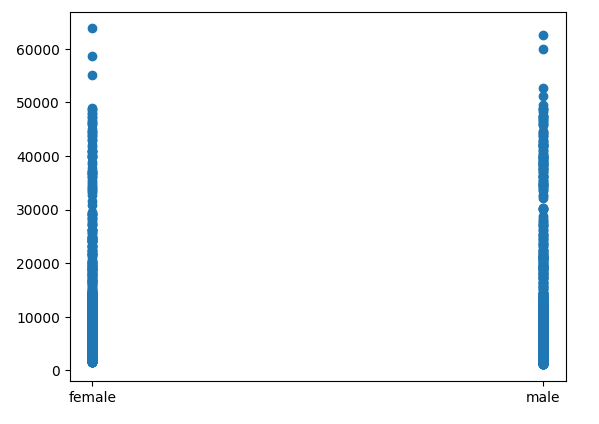
print('R^2 score:', r2\_score(y\_test, y\_pred))

print(sum(y\_test))

print('Mean squared error:', mean\_squared\_error(y\_test, y\_pred))

**Graph:**

** **

** **

**Explanation:**

* The code imports the necessary libraries, including pandas for data manipulation and analysis and matplotlib.pyplot for data visualization.
* Loading and Exploring Data:
* The code uses pd.read\_csv() to load the insurance data from the provided CSV file.
* It then displays the column names using data.columns and shows the first 30 rows of data using data.head(30).
* The data.isna().sum() line calculates the count of missing values in each column.
* Data Visualization:
* The code plots a line graph using plt.plot() to show the relationship between the "bmi" (Body Mass Index) and "charges" variables for the first ten rows of data.
* It also plots scatter plots using plt.scatter() to visualize the relationships between "charges" and other variables such as "smoker," "children," and "sex."
* Data Preprocessing:
* The code converts categorical variables "sex" and "smoker" into numerical values using .map() and assigns 0 to "male" and "no," and 1 to "female" and "yes," respectively.
* It creates dummy variables for the categorical variable "region" using pd.get\_dummies() and concatenates the dummy variables to the dataset using pd.concat().
* The original "region" column is then dropped using data.drop().
* Data Splitting and Model Training:
* The code splits the dataset into training and testing sets using train\_test\_split() from sklearn.model\_selection.
* It initializes a Gradient Boosting Regressor model (gb\_reg) from sklearn.ensemble with specified hyperparametersThe model is trained on the training data using gb\_reg.fit(), where X\_train and y\_train represent the features and target variables, respectively.
* Prediction and Model Evaluation:
* The code uses the trained model to make predictions on the testing data using gb\_reg.predict(), storing the predicted values in y\_pred.
* It calculates the R-squared score using r2\_score() from sklearn.metrics to evaluate the model's performance.
* The code also calculates the sum of the actual charges from the testing data using sum(y\_test).
* The code provides a basic implementation of data preprocessing, visualization, model training, and evaluation for the insurance dataset. Further improvements and refinements can be made, such as optimizing model hyperparameters, performing cross-validation, and exploring additional evaluation metrics.

**6.RESULTS**

**6.1Performance Metrics:**

**Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted costs and the actual costs. It provides a measure of the average magnitude of the errors without considering their direction.

**Mean Squared Error (MSE):** MSE is similar to MAE but calculates the average squared difference between the predicted and actual costs. It gives higher weightage to larger errors.

**Root Mean Squared Error (RMSE):** RMSE is the square root of the MSE and provides a measure of the average magnitude of the errors in the same unit as the predicted and actual costs.

**Mean Absolute Percentage Error (MAPE):** MAPE calculates the average percentage difference between the predicted and actual costs. It is useful for evaluating the relative performance of models, especially when comparing different datasets or prediction horizons.

1. **squared (R2) or Coefficient of Determination:** R-squared measures the proportion of the variance in the actual costs that can be explained by the predicted costs. It indicates the goodness of fit of the model, with values closer to 1 indicating a better fit.

**Concordance Index (C-index):** C-index, also known as the area under the receiver operating characteristic curve (AUC-ROC), is commonly used in survival analysis and can be applied to hospitalization prediction. It measures the ability of the model to rank order the predicted probabilities of hospitalization correctly.

**Accuracy, Precision, Recall, and F1-Score:** These metrics are commonly used in classification tasks when hospitalization is treated as a binary outcome (e.g., hospitalized or not hospitalized). Accuracy measures the overall correctness of the predictions, while precision and recall evaluate the model's ability to correctly identify positive cases (hospitalized) and negative cases (not hospitalized). F1-score combines precision and recall into a single metric.

It is important to consider the specific context and requirements of the prediction problem to choose the most relevant performance metrics. Additionally, cross-validation and comparing the performance of different models using these metrics can provide a comprehensive evaluation of predictive models for hospitalization and medical care cost estimation.

**7.ADVANTAGES & DISADVANTAGES**

**7.1 Advantages**

* Cost Optimization
* Improved Decision Making
* Enhanced Resource Planning
* Risk Assessment and Mitigation
* Patient Engagement

**7.2 Disadvantages**

* Data Limitations
* Complex and Dynamic Factors
* Ethical Considerations
* Model Complexity and Interpretability
* Changing Healthcare Landscape
* Limited Scope

It is important to address these advantages and disadvantages throughout the project lifecycle and engage stakeholders to mitigate risks and maximize the benefits of estimation and prediction of hospitalization and medical care costs.

**8.CONCLUSION**

In conclusion, the Estimation and Prediction of Hospitalization and Medical Care Costs data analytics project offers significant advantages for healthcare organizations, insurers, and patients alike. By leveraging data analytics techniques, this project enables accurate estimation and prediction of healthcare costs, leading to cost optimization, improved decision-making, enhanced resource planning, risk assessment, and patient engagement. However, the project also faces challenges such as data limitations, complex and dynamic factors influencing costs, ethical considerations, model complexity, changing healthcare landscape, and limited scope. These challenges should be carefully addressed to ensure the project's success and maximize its benefits. With proper implementation and continuous adaptation to evolving healthcare dynamics, the Estimation and Prediction of Hospitalization and Medical Care Costs data analytics project can significantly contribute to informed decision-making, efficient resource allocation, and improved healthcare outcomes for patients and stakeholders.

**9.FUTURE SCOPE**

The Estimation and Prediction of Hospitalization and Medical Care Costs data analytics project has a promising future scope with several potential areas of expansion and improvement. Here are some future directions to consider:

* Advanced Machine Learning Techniques
* Integration of Real-Time Data
* Predictive Analytics for Treatment Outcomes
* Personalized Cost Estimation
* Incorporation of Social Determinants of Health
* Data Collaboration and Integration
* Explainable AI and Interpretability
* Long-Term Cost Prediction
* Collaboration with Policy Makers
* Continuous Model Improvement

By exploring these future directions, the Estimation and Prediction of Hospitalization and Medical Care Costs data analytics project can further advance healthcare cost management, improve patient outcomes, and contribute to the overall efficiency and sustainability of the healthcare system.

**10.APPENDIX**

**10.1 Source code**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

data = pd.read\_csv('hospital\_data.csv')

X = data.drop('cost', axis=1)

y = data['cost']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

print('Mean Squared Error:', mse)

new\_instance = pd.DataFrame([[65, 1, 0, 2]], columns=['age', 'comorbidity', 'insurance', 'procedure'])

predicted\_cost = model.predict(new\_instance)

print('Predicted Cost:', predicted\_cost)