**Mini Project Synopsis**

**1. Project Title:**  
 ‘Healthcare Cost Estimation System’

**2. Problem Statement:**  
This project aims to develop a machine learning model that can estimate the cost of healthcare based on various features such as patient age, diagnosis, and type of treatment.

**3. Objective:**  
The model will be trained on a healthcare dataset, which contains historical data on patient characteristics, medical conditions, and corresponding healthcare costs.

**4. Proposed Solution:**  
We will be using a linear regression model to create this project.

**5. Expected Outcome:**  
The outcome of this project will be a machine learning model that can accurately predict the cost of healthcare based on patient characteristics, diagnosis, and type of treatment.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student’s signature Supervisor’s signature

The provided code performs a comprehensive analysis of health insurance charges using a dataset. Here's a breakdown of the key steps and components of the code:

**1. Data Preparation**

* **Import Libraries**: The code imports necessary libraries such as **pandas**, **numpy**, **seaborn**, **matplotlib**, **statsmodels**, and **sklearn**.
* **Load Data**: The dataset is loaded from a CSV file named **insurance.csv**.
* **Transform Charges**: The **charges** column is transformed using the natural logarithm to stabilize variance and normalize the distribution.

**2. Data Exploration**

* **Convert Columns to Numeric**: The columns **age**, **bmi**, and **children** are explicitly converted to numeric types.
* **Visualizations**: Several plots are created to explore relationships between **charges** and other variables:
  + Scatter plots for **age** vs. **charges** and **bmi** vs. **charges**.
  + Box plots for **charges** by **sex**, **children**, **smoker**, and **region**.

**3. Feature Engineering**

* **One-Hot Encoding**: Categorical variables (**sex**, **smoker**, **region**) are converted into dummy/indicator variables using one-hot encoding.

**4. Model Training**

* **Train-Test Split**: The dataset is split into training and testing sets (80% train, 20% test).
* **Model 0**: An Ordinary Least Squares (OLS) regression model is fitted using all features. The model summary is printed, and predictions are made on the test set. The Root Mean Squared Error (RMSE) is calculated.
* **Model 1**: A second OLS model is fitted using only significant features identified from the first model. The summary and RMSE are printed.

**5. Model Evaluation**

* **Prediction vs. Actual**: A scatter plot is created to visualize the relationship between predicted and actual charges.
* **Residual Analysis**: Residuals (differences between actual and predicted values) are plotted against predicted values, and a histogram of residuals is displayed to check for normality.

**6. Prediction Function**

* A function **predict\_charges** is defined to make predictions for new data points based on the fitted model. It takes parameters for **age**, **bmi**, **children**, **smoker**, and **region**.
* Predictions are made for three individuals (Bob, Lisa, and John) using the function, and the results are printed.

**Example Output**

The output will include:

* Summary statistics of the dataset.
* RMSE values for both models.
* Predicted health care charges for Bob, Lisa, and John.

**Important Notes**

* Ensure that the dataset **insurance.csv** is available in the specified path for the code to run successfully.
* The code assumes that the necessary libraries are installed in your Python environment.
* The transformation of the **charges** variable may affect the interpretation of the model coefficients, as they are in log scale.

**Conclusion**

This code provides a solid framework for analyzing health insurance charges, building predictive models, and evaluating their performance. It can be further enhanced by exploring additional features, tuning model parameters, or trying different modeling techniques.

Report data-

**Chapter 1: Introduction and Problem Statement**

1.1 **Introduction**  
The healthcare industry is increasingly relying on data-driven approaches to predict costs and improve patient care. This project focuses on predicting health care charges based on various factors such as age, BMI, number of children, smoking status, and region. The analysis utilizes a dataset containing insurance charges to build a predictive model.

1.2 **Problem Statement**  
The problem statement for the present work can be stated as follows:

* To develop a predictive model that estimates health care charges based on individual characteristics and lifestyle choices.

**Chapter 2: Methodology**

* **Data Collection:** The dataset used for this project is the "insurance.csv" file, which contains information about individuals' health insurance charges.
* **Data Preprocessing:**
  + Log transformation of the 'charges' column to normalize the distribution.
  + Conversion of specific columns to numeric types.
  + Creation of dummy variables for categorical features.
* **Model Development:**
  + Split the dataset into training and testing sets.
  + Build an Ordinary Least Squares (OLS) regression model.
  + Evaluate model performance using Root Mean Squared Error (RMSE).

**Chapter 3: Project Work Carried Out**

* **Data Visualization:**
  + Scatter plots to visualize the correlation between age, BMI, and charges.
  + Box plots to analyze charges based on sex, number of children, smoking status, and region.
* **Model Fitting:**
  + Initial model fitted with all features.
  + A second model fitted with only significant features identified from the first model.
* **Model Evaluation:**
  + RMSE calculated for both models to assess prediction accuracy.

**Chapter 4: Results and Discussion**

* **Model Summary:**
  + The first model summary indicates the overall fit and significance of predictors.
  + The second model summary highlights the significant features impacting health care charges.
* **RMSE Results:**
  + RMSE for the first model: [Insert RMSE value]
  + RMSE for the second model: [Insert RMSE value]
* **Visualizations:**
  + Scatter plot comparing predicted vs. actual charges.
  + Residuals plot to check for homoscedasticity.
  + Histogram of residuals to assess normality.

**Chapter 5: Conclusion and Future Work**

* **Conclusion:**  
  The predictive model successfully estimates health care charges based on individual characteristics. The analysis demonstrates the importance of factors such as age, BMI, and smoking status in determining charges.
* **Future Work:**  
  Future enhancements could include:
  + Incorporating additional features such as medical history.
  + Exploring advanced machine learning techniques for improved accuracy.

**Guide Interaction Form**

[To be filled out during the project]

**References**

1. [Insert relevant references used in the project]
2. [Insert relevant references used in the project]
3. [Insert relevant references used in the project]