

**Fall Semester: 2024-25**

**Lab Experiments**

**CSE2001-Object Oriented Programming using C++**

**Slot – B14+C14+D14+E14+F14**

**Faculty Name – Mr. Venkateswara Rao T**

Submitted By:

Ananya Kumari 23BCE10659

Priya Singh 23BCE11177

Suyyash Arora 23BSA10033

ACKNOWLEDGEMENT

First and foremost we would like to thank ALMIGHTY who has provided us the strength to do justice to our work and contribute our best to it We wish to express our deep sense of gratitude from the bottom of our heart to our guide Mr. Venkateswara Rao T. Professor of Computer Science and Engineering, for his motivating discussion, overwhelming suggestions, ingenious encouragement, invaluable supervision and exemplary guidance throughout this project work . As a final word, we would like to thank each and every individual who have been a source of support and encouragement and helped us to achieve our goal and complete our project work successfully

ABSTRACT

Medical costs are one of the most common reoccurring expenses in a person's life. It is general known that a person's lifestyle and numerous physical factors determine the diseases or disorders they may get, and that these conditions determine medical expenses. According to several research, there are several significant reasons that lead to greater expenditures. smoking, age, and BMI are all factors in personal medical care. The goal of this study is to examine and identify a link between personal medical costs and other characteristics. Then, by generating linear regression models and comparing them using ANOVA, we use the significant traits as predictors to forecast medical expenditures. In our research, we discovered that smoking, age, and a higher BMI all have a significant connection with higher medical expenditures, showing that they are key contributors to the charges, and that the regression can predict the charges with more than 75% accuracy. According to the World Health Organization, personal medical and healthcare spending is growing faster than the global economy This rise in spending has been related to a variety of factors, the most prominent of which are smoking, ageing, and higher BMI. Using insurance data from diverse persons with variables such as smoking, age, number of children, area, and BMI, we hope to uncover a link between medical expenditures and other parameters.

INTRODUCTION

OVERVIEW

The expense of health care is rising every day. There is a need to forecast health costs as the number of novel viruses infecting humans grows. This form of forecasting aids governments in making health-related decisions. People are also aware of the significance of health-care spending. Machine Learning is a field that touches all aspect of life. Machine learning models are also used in the health-care system for a variety of health-related applications. We conducted a predicate analysis on medical health insurance expenses in this study. We create a model to forecast a person's medical insurance costs depending on gender. The dataset comes from Kaggle and comprises 1338 rows of data with the following attributes: age, gender, smoker, BMI, children, region, and insurance charges. Medical information and expenditures billed by health insurance companies are included in the data. To forecast medical expenses, we used a variety of regression techniques on this dataset. The Python programming language was utilized to implement the project. In 2015, the absolute consumption of medical services as a proportion of GDP was 3.89 percent, according to the World Bank. Legal medical usage amounts for barely 1% of GDP, down from 3.89 percent in 2015, and cash-based medical use makes for 65.06 percent of current medical use. Over the previous few decades, advances in clinical innovation have made it feasible to cure ailments. It deals with a subject that was formerly supposed to be fatal. In any event, the expense of her therapy is so high that it is unaffordable for someone in the white-collar class. A 5,000 rupee floater plan for yourself, your spouse, and your children is anticipated to cost 10,000 to 17,000 rupees per year, while a 5,000 rupee sickness plan will cost 4,000 rupees for multiple people.. In any case, the expense of their therapy is so high that it is practically impossible for a white-collar worker to pay it. A 5,000 rupee floater plan for a family

will cost between 10,000 and 17,000 rupees per year, whereas a 5,000 rupee health insurance plan for multiple years will cost 4,000 rupees. It ranges in price from a few rupees to 7,000 rupees. It'll take a year. According to the calculations. It then generates linear regression models and

compares them using ANOVA to employ an important function as a predictor for forecasting medical expenditures. Our research discovered that smoking, being older, and having a higher BMI were all linked to higher medical expenditures. This demonstrates that these are the primary sources of cost, and that regression can accurately estimate expenses with a 75% accuracy rate

Problem Definition and Scenarios

In the current world, the medical expenses plays the vital role in life. sometimes we need to pay huge amount for treatments. We are going to build a system, which predicts the medical insurance amount from the company based on certain parameters and which also provides cost of treatment for specific disease. With help of machine learning we are going to build a system, which takes inputs from the user and predicts the medical treatment cost of the disease and insurance claim amount. for this we are using linear regression. Linear regression: Predict the response variable's result using some explanatory factors

Literature:

Introduction to machine learning

Machine learning is a technique that allows computer to learn from past data and anticipate fresh samples. Machine Learning models may be used in any sector. Medical records are likewise not exempt from machine learning. For numerous years, the medical industry has used models in various settings. Many of the studies used machine learning approaches to forecast medical costs B. Nithya [1] et.al In Predictive Analytics in Health Care, machine learning models were used. For predictive analysis, they used a variety of supervised and unsupervised models. They also claimed that machine learning tools and techniques are crucial in health-care sectors, and that they are exclusively employed in the detection and prognosis of various malignancies. Ahuja Tike[2] et.al applied hierarchical decision trees for the medical price prediction systems. Their experiments showed that the price prediction system achieves high accuracy. Moran et al. [3] utilized linear regression techniques to anticipate Intensive Care Unit (ICU) expenses and utilize understanding socioeconomics, DRG (Diagnostic Related Group), length of stay in the clinic, and a couple of others as highlights. Gregory [4] et.al applied various regression models for analyzing medical costs in the health care system. They mainly concentrated on reducing the bias in the cost estimates to achieve good results. Dimitris Bertsimas[5] et.al applied different data mining techniques which provided an accurate prediction of medical costs and represent a powerful tool for the prediction of healthcare costs.

Medical Expense Prediction System using Machine learning Techniques and Intelligent Fuzzy Approach (2020, H. Chen Jonathan, M. Asch Steven) Prediction isn't a new concept in medicine. Clinical predictions based on data are becoming commonplace in medicine., ranging Risk categorization of patients in the critical care unit ranges from risk scores to anticoagulant treatment (CHADS2) and cholesterol medicine usage (ASCVD) (APACHE). You may easily develop prediction models for hundreds of similar clinical questions using clinical data sources and contemporary machine learning. These approaches might be used for everything from sepsis early warning systems to superhuman diagnostic imaging. The real data source, on the other hand, has an issue. Unlike traditional techniques, which rely on data from cohorts that have been thoroughly prepared to prevent bias, new data sources are sometimes unstructured due to the fact that they were developed for various purposes (clinical care, billing, etc.). Patient self-selection, indication misunderstanding, and inconsistent outcome data can all contribute to unintentional biases and even racist programming in machine prediction. As a result of this understanding, discussing the potential of data analysis to aid medical decision-making isn't just wishful thinking

**Objective of the Project work:**

The objective of proposed work is to predict the risk for medical insurance and identify those patients most at risk of being re-admitted. It means that patients can have greater support after discharged from hospital. There are some risks types which is required to be predicted for a patient.

* Cost & Utilization: The expected cost for a member in the next X years, Predicting emergency department visits. The likelihood of a hospital re-admission for recently discharged patients
* Clinical: The likelihood of a person suffering an acute event, such as an acute myocardial infection. The expected disease progression (cost, visit, and/or clinical severity) of someone with a mental illness. A person’s expected risk of mortality.
* Program-focused: A patient’s likelihood to engage in a mobile health program. An illustration of how healthcare providers can take advantage of machine learning is being able to predict hospital re-admission for chronically ill patients. While the healthcare sector is being transformed by the ability to record massive amounts of information about individual patients, the enormous volume of data being collected is impossible for human beings to analyses. Machine learning provides a way to automatically find patterns and reason about data, which enables healthcare professionals to move to personalized care known as precision medicine. There are many possibilities for how machine learning can be used in healthcare, and all of them depend on having sufficient data and permission to use it. Previously, alerts and recommendations for a medical practice have been developed based on external studies, and hard-coded into their software. However, that can limit the accuracy of that data because they might be from different populations and environments.

**Software and Hardware requirement:**

**Software:**

1. **Jupyter Notebook(for line by line execution of code)**
2. **Python**
3. **Required modules**

**Hardware:**

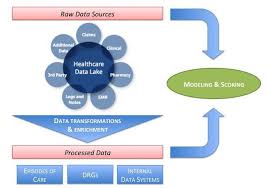
1. **Working computer with at least 4gb ram**

**Methodology**

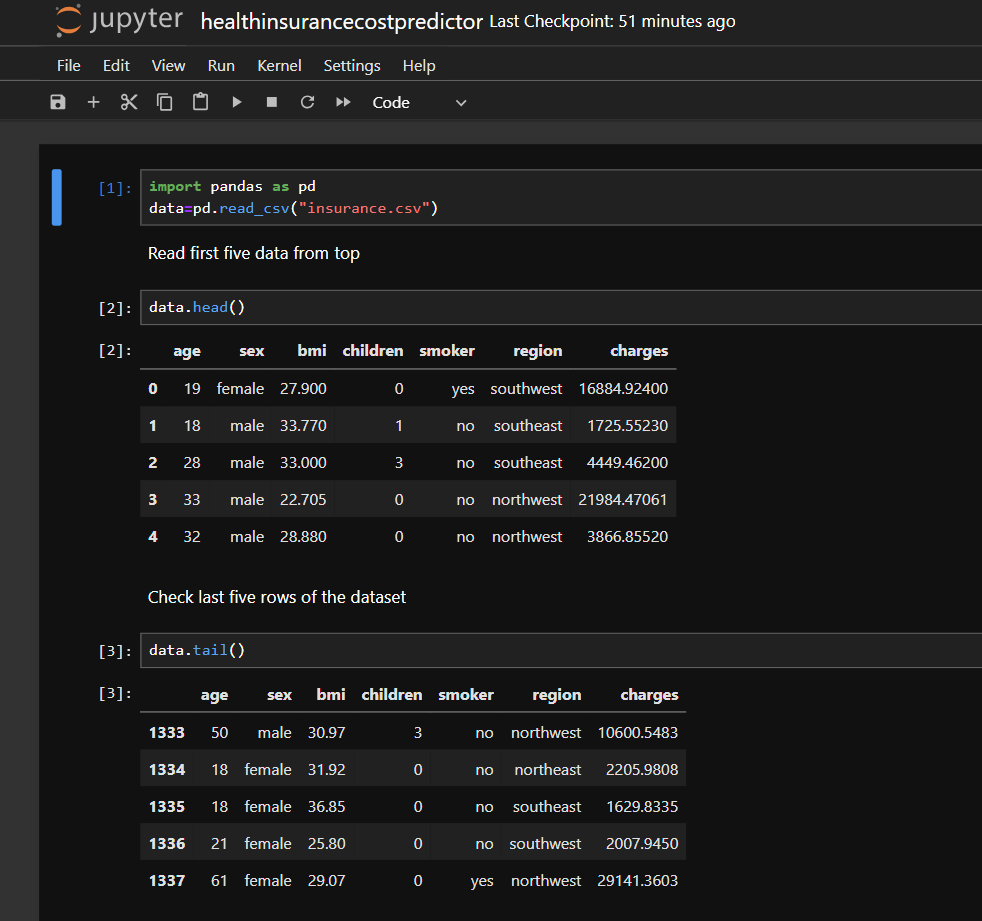
**Step 1: Data collection**: This will involve collection of student feedback in the form of structured data like the grades, enrollment data, progression rates as well as unstructured data like student opinions expressed through surveys, web blogs, twitter, Facebook etc.

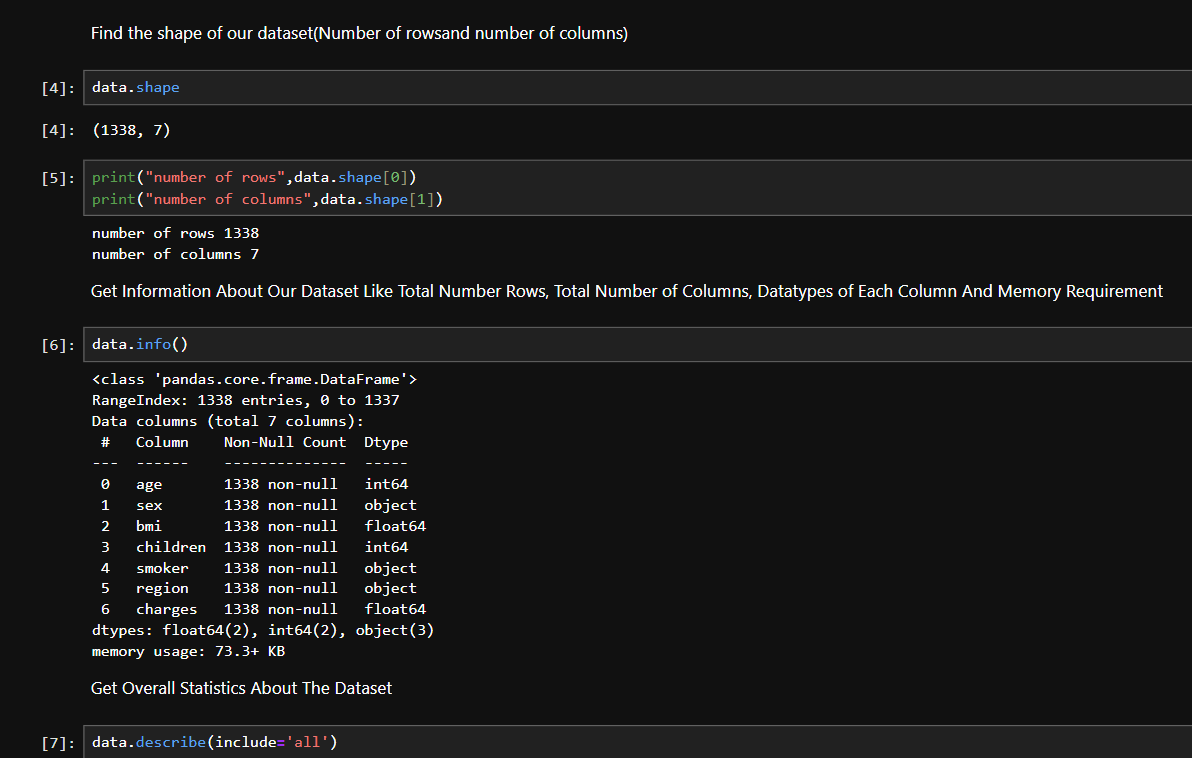
**Step 2: Data Preprocessing**: In this phase, the data is prepared for the analysis purpose which contains relevant information. Pre-processing and cleaning of data are one of the most important tasks that must be one before dataset can be used for machine learning. The real-world data is noisy, incomplete and inconsistent. So, it is required to be cleaned.

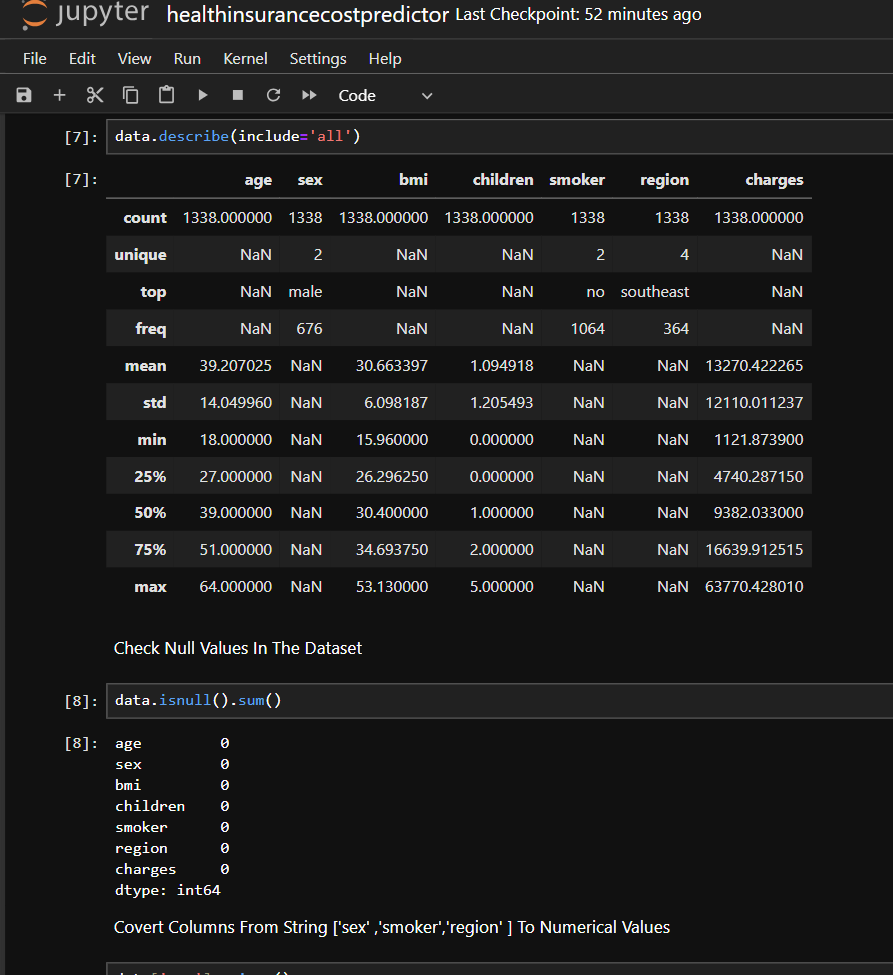
**Step 3: Extraction of Feature Set/Training Dat**a Feature set or training data can be prepared from the cleaned data by using any of the available techniques like bag of words, -gram, N-gram, POS, TOS tagging etc. The training data can also be prepared by providing them labels and then divide it into two classes like positive class and negative class. The feature sets and training set that has obtained by using any of the above methods will be used for the implementation of machine learning algorithms.

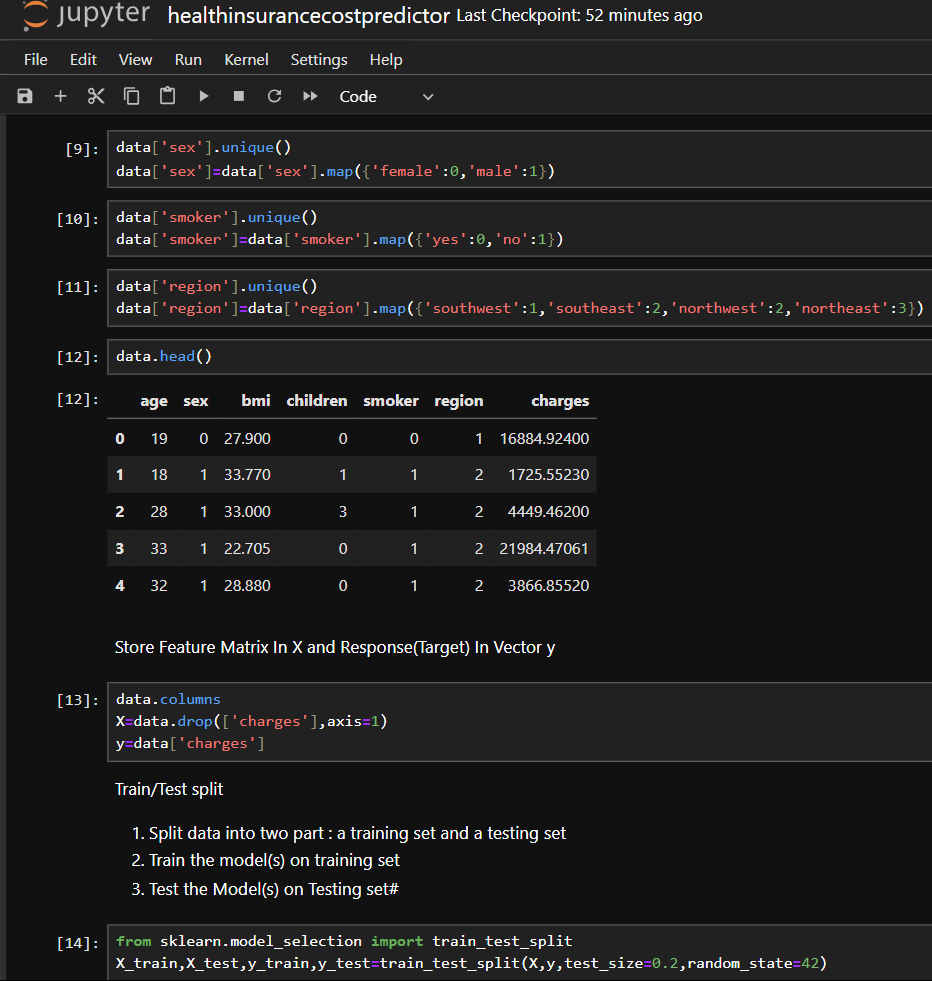
**Step 4: Implementation of Machine Learning Algorithm on Feature Set/Training Dat**a Classification: To determine a label or category – it is either one thing or another. We train the model using a set of labelled data. As an example, we want to predict if a person’s mole is cancerous or not, so we create a model using a data set of mole scans from 1000 patients that a doctor has already examined to determine whether they show cancer or not. We also feed the model a whole bunch of other data such as a patient’s age, gender, ethnicity, and place of residence. Then create a model which will enable us to present a new mole scan & decide if it depict cancer or not. Regression: A Regression model is created when we want to find out a number – for example how many days before a patient discharged from hospital with a chronic condition such as diabetes will return. Step 5: Testing of Data Testing of data is done based on training model which is classified using supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in context of the given data.

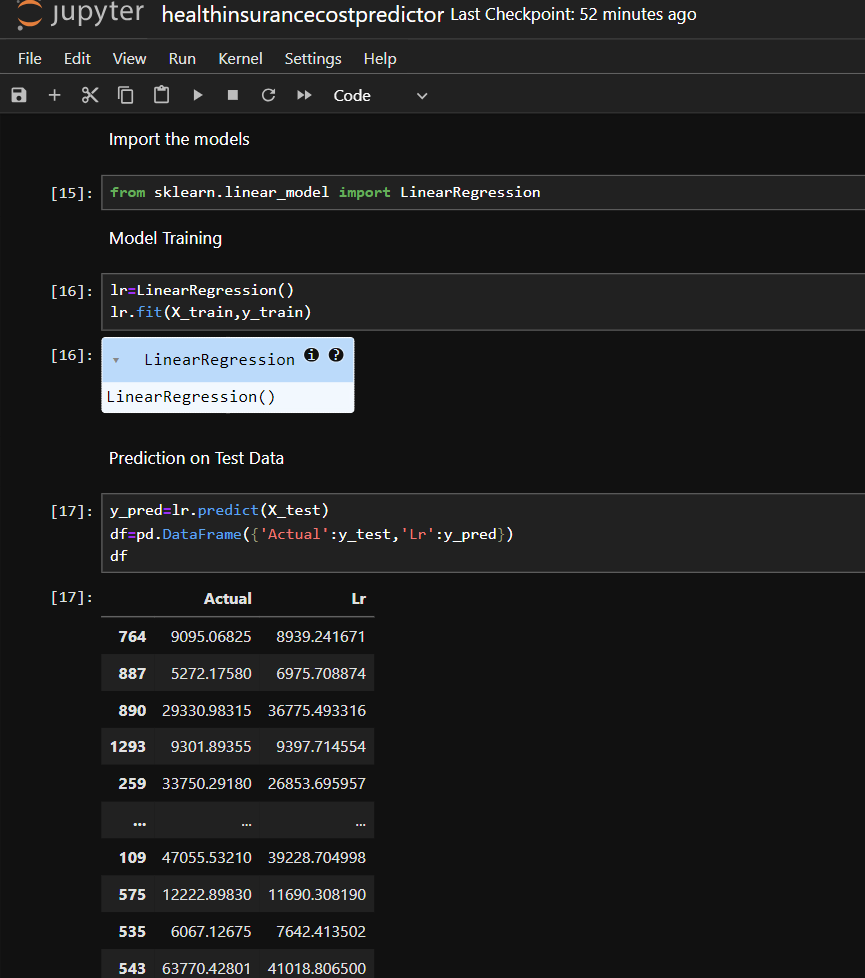
**Implementation:**

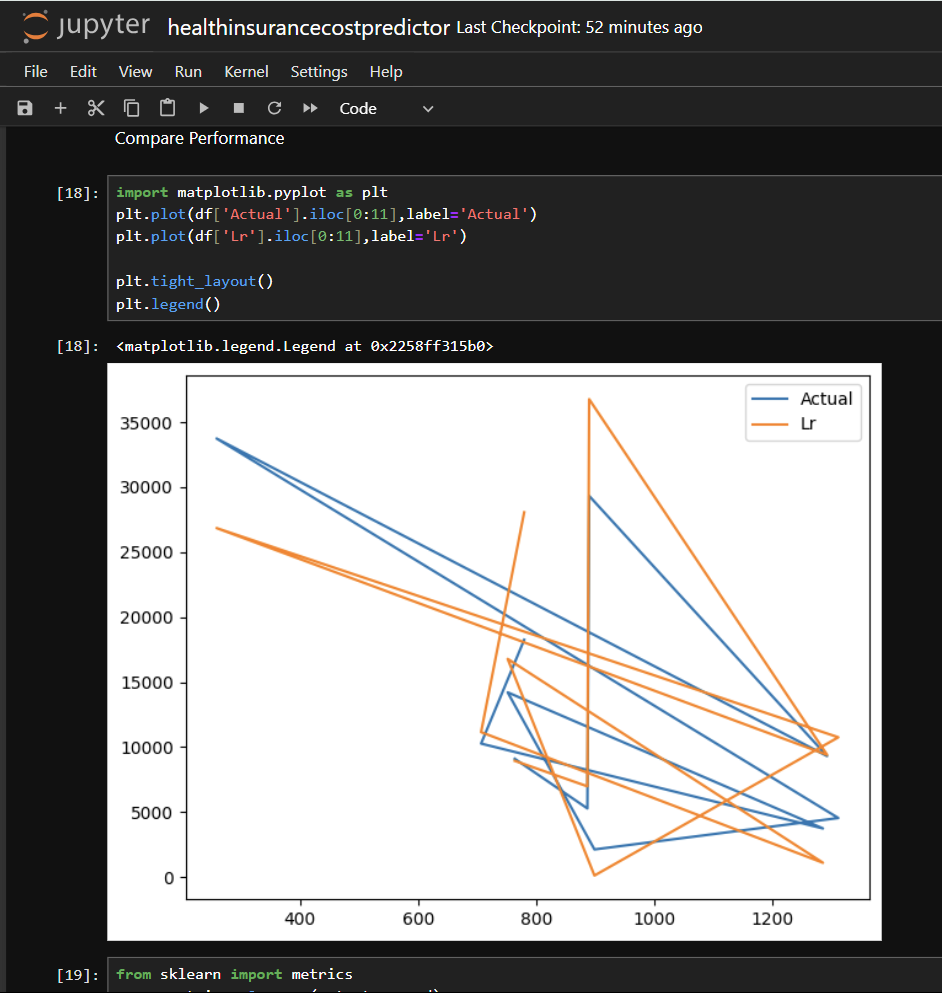
****

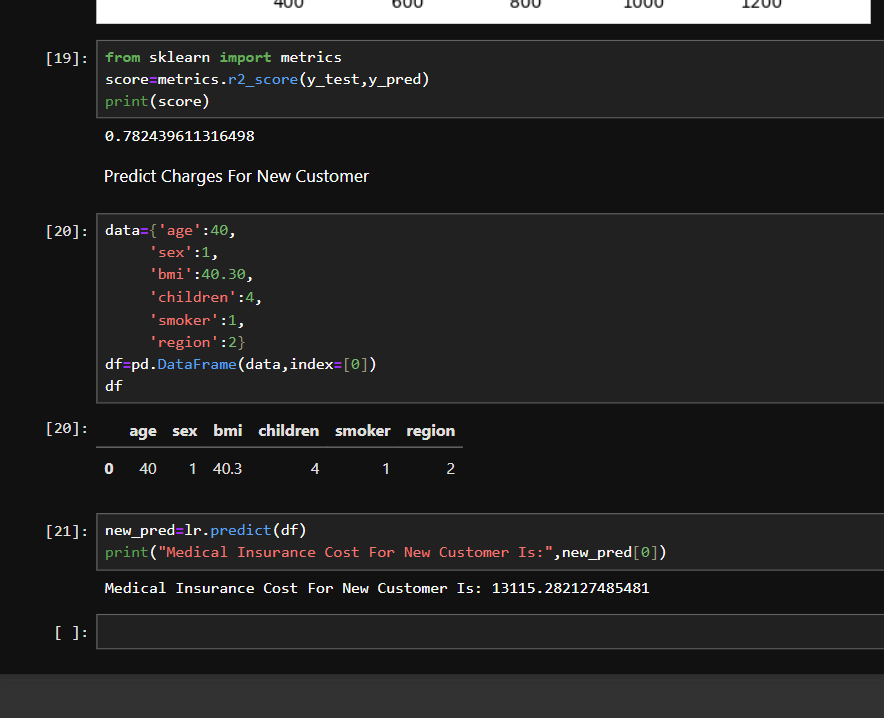
****











Results:

The results of a health insurance cost predictor project using machine learning can vary depending on several factors, including the quality of your data, the features you use, and the algorithms you implement. Here’s a general overview of what you might expect:

**Model Performance Metrics**:

**Accuracy**: Measures how often the model’s predictions are correct. For regression tasks (predicting costs), you might look at metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

**R-Squared (R²)**: Indicates how well the model explains the variability of the target variable. A higher R² value means a better fit.

**Feature Importance**:

You’ll often analyze which features (e.g., age, BMI, smoking status, etc.) are most influential in predicting insurance costs. This helps in understanding what factors most affect the cost predictions.

**Model Comparisons**:

You might compare different models (e.g., linear regression, decision trees, random forests, gradient boosting) to determine which one provides the best performance.

**Error Analysis**:

Analyzing the residuals (the difference between predicted and actual costs) helps to understand any patterns the model may be missing.

**Visualizations**:

Graphs and plots, such as predicted vs. actual values, feature importance plots, and residual plots, can provide insights into model performance and areas for improvement.

**Implementation**:

Once a model is selected, it’s important to validate it on new data to ensure its robustness and reliability before deployment.

Limitations of this project:

Assumption of Linearity:

Linear regression assumes a linear relationship between the independent variables (age, sex, BMI, children, smoking habits, region) and the dependent variable (annual medical expenditure). Real-world data might exhibit more complex, non-linear relationships which a linear model cannot capture.

Sensitivity to Outliers:

Linear regression is sensitive to outliers. Extreme values in the data can disproportionately influence the model, leading to inaccurate predictions.

Multicollinearity:

When predictor variables are highly correlated, it can cause multicollinearity, which can make it difficult to determine the individual effect of each predictor on the dependent variable and can lead to less reliable coefficient estimates.

Limited Complexity:

Linear regression cannot capture complex interactions between features. For example, the interaction effect between smoking and age on medical costs may not be effectively captured by a simple linear model.

Overfitting and Underfitting:

If the model is too simple, it may underfit the data, failing to capture the underlying trend. Conversely, if too many features or high-degree polynomial terms are added, it may overfit, capturing noise rather than the actual trend.

Assumption of Homoscedasticity:

Linear regression assumes constant variance of the errors (homoscedasticity). If the variance of errors differs across levels of an independent variable (heteroscedasticity), the model’s predictions may be inefficient and biased.

Normality of Errors:

The model assumes that the residuals (errors) are normally distributed. If this assumption is violated, it can affect the reliability of the confidence intervals and hypothesis tests.

Data Quality and Quantity:

The accuracy of predictions heavily depends on the quality and quantity of the training data. Incomplete or biased data can lead to inaccurate predictions.

Feature Limitations:

Important predictors might be missing from the model. For example, factors like genetic predisposition, lifestyle choices, and specific medical history are not included, which can significantly affect medical costs.

Interpretability and Generalization:

While linear regression models are generally easy to interpret, the simplicity might come at the cost of not capturing complex patterns in the data. Additionally, the model’s performance might degrade when applied to new, unseen data if the training data does not represent the broader population well.

Addressing Limitations

To address these limitations, consider the following approaches:

Use more advanced regression techniques (e.g., polynomial regression, regularization methods like Ridge or Lasso).

Explore non-linear models (e.g., decision trees, random forests, gradient boosting, neural networks).

Perform thorough data preprocessing, including outlier treatment, multicollinearity checks, and ensuring data quality.

Use cross-validation techniques to assess model performance and prevent overfitting.

Include a diverse set of relevant features and explore feature engineering to capture complex interactions.

Discussion:

At the end of a report on a health insurance cost predictor project using machine learning, you should aim to provide a comprehensive summary and reflect on the key aspects of the project. Here’s a structured outline for concluding the report:

**1. Summary of Findings**

* **Project Overview**: Recap the purpose and objectives of the project. Explain how the machine learning model was designed to predict health insurance costs.
* **Key Results**: Summarize the main findings, including the performance of the model, important features, and any notable patterns discovered during analysis.

**2. Model Performance**

* **Performance Metrics**: Reiterate the key performance metrics (e.g., MAE, MSE, RMSE, R²) and discuss the implications of these results. Highlight how well the model met the initial goals.
* **Model Comparison**: Briefly review the performance of different models tested and justify why the final model was chosen.

**3. Insights and Implications**

* **Feature Insights**: Highlight the most important features affecting insurance costs and their potential impact on decision-making.
* **Business Implications**: Discuss how the model’s predictions could be used in practice by insurance companies, such as for pricing strategies, risk assessment, or customer segmentation.

**4. Challenges and Limitations**

* **Data Issues**: Reflect on any challenges faced with data quality, such as missing values, biases, or outliers.
* **Model Limitations**: Address any limitations of the model, including areas where it may have underperformed or any assumptions that may affect its accuracy.

**5. Recommendations for Future Work**

* **Model Improvement**: Suggest ways to enhance the model, such as incorporating additional features, exploring more advanced algorithms, or refining hyperparameters.
* **Additional Research**: Recommend further research or analysis that could provide additional insights or improve the model’s accuracy and applicability.

**6. Ethical and Regulatory Considerations**

* **Data Privacy**: Reiterate how data privacy and security were handled during the project.
* **Bias and Fairness**: Discuss any steps taken to ensure the model is fair and unbiased, and suggest how to address any potential issues in future iterations.

Conclusion

In conclusion, our health insurance cost predictor project using a linear regression machine learning model represents a significant step towards automating and enhancing the accuracy of medical expenditure estimation. By leveraging demographic and lifestyle factors such as age, sex, BMI, number of children, smoking habits, and region of residence, our model provides a streamlined method for predicting annual health insurance costs.

**References:**

<https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset>

<https://chatgpt.com/c/f8b110f2-a9f4-4592-99de-d5ea125daa6f>

<https://github.com/AlishahShaikh/Health-Insurance-Cost-Prediction->