Introduction

Medical expenses are any costs incurred in the prevention or treatment of injury or disease. Medical expenses include health and dental [insurance premiums](https://www.investopedia.com/terms/i/insurance-premium.asp), doctor and hospital visits, [co-pays](https://www.investopedia.com/terms/c/copay.asp), prescription and over-the-counter drugs, glasses and contacts, crutches, and wheelchairs, to name a few.

Medical cost insurance plays a crucial role in healthcare , providing individuals and families with financial protection against the high costs of medical treatment and healthcare services. Its primary objective is to address the financial burden associated with unexpected medical emergencies, routine healthcare services, and chronic illnesses.

Accurately predicting medical costs is a complex task influenced by various factors, including demographic characteristics, pre-existing health conditions, lifestyle choices, and regional healthcare policies.

The advent of machine learning and predictive analytics offers promising opportunities to enhance the accuracy and efficiency of medical cost insurance prediction. By leveraging advanced algorithms and large-scale healthcare datasets, insurers, healthcare providers, and policymakers can gain valuable insights into the drivers of medical costs and develop data-driven strategies for risk assessment, pricing, and resource allocation.

Methodology

1. \*\*Data Collection\*\*

The medical cost insurance prediction dataset from Kaggle, a popular platform for sharing and discovering datasets and data science projects. The dataset contains information on medical costs incurred by individuals and includes various features

- The specific dataset used for this project is publicly available and can be accessed by registered users on the Kaggle platform.

- The dataset contains enough records to enable robust model training and evaluation, with features representing a wide range of factors potentially influencing medical costs

\*\*Dataset Description:\*\*

- The dataset comprises records of medical cost insurance claims filed by individuals.

- Each record includes the following features:

- \*\*Age\*\*: Age of the individual in years.

- \*\*Sex\*\*: Gender of the individual (male or female).

- \*\*BMI\*\*: Body mass index, a measure of body fat based on height and weight.

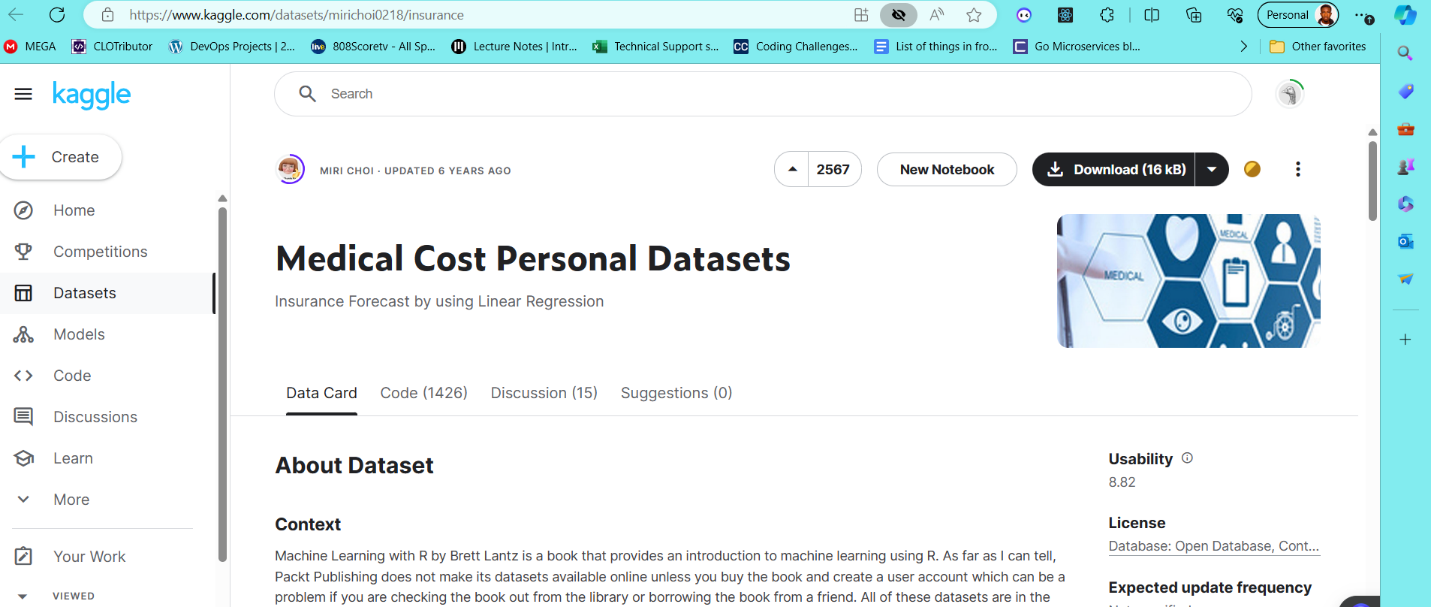
- \*\*Children\*\*: Number of children/dependents covered by the insurance plan.

- \*\*Smoker\*\*: Whether the individual is a smoker or not (yes or no).

- \*\*Region\*\*: Geographic region of residence (e.g., northeast, southeast, southwest, northwest).

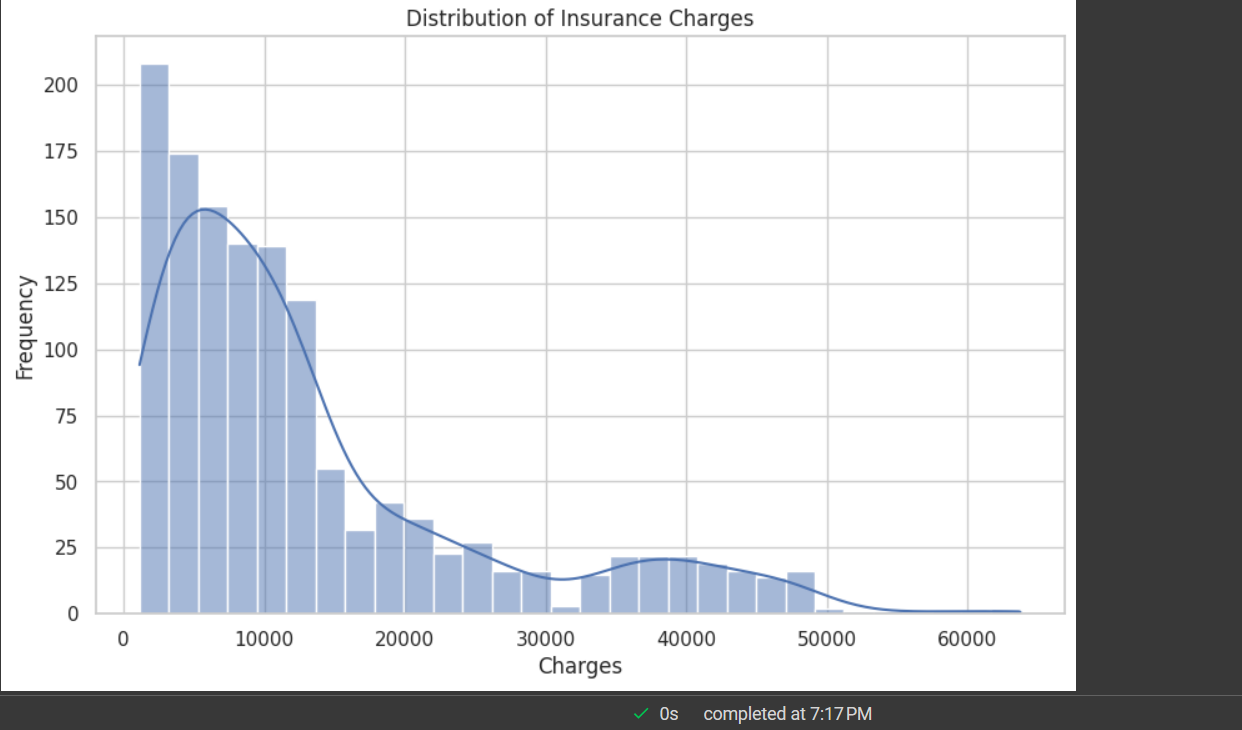
- \*\*Charges\*\*: Medical costs incurred by the individual.

..



1. Data Preprocessing

The medical insurance cost dataset was a very clean dataset with little or no preprocessing to do, there were no missing fields cross rows and columns so it was easy to handle initially. As a result of this, insights into the data via visualization was the focus during the preliminary data handling phase. Some of the visualizations can be seen below between features in the dataset.

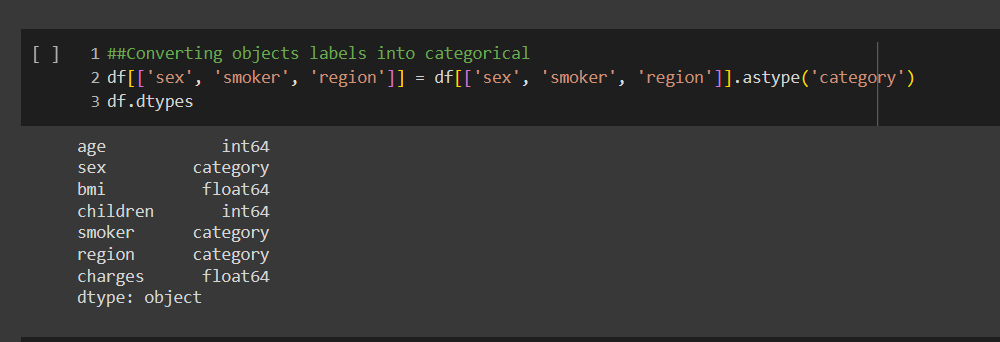


A screenshot of a graph

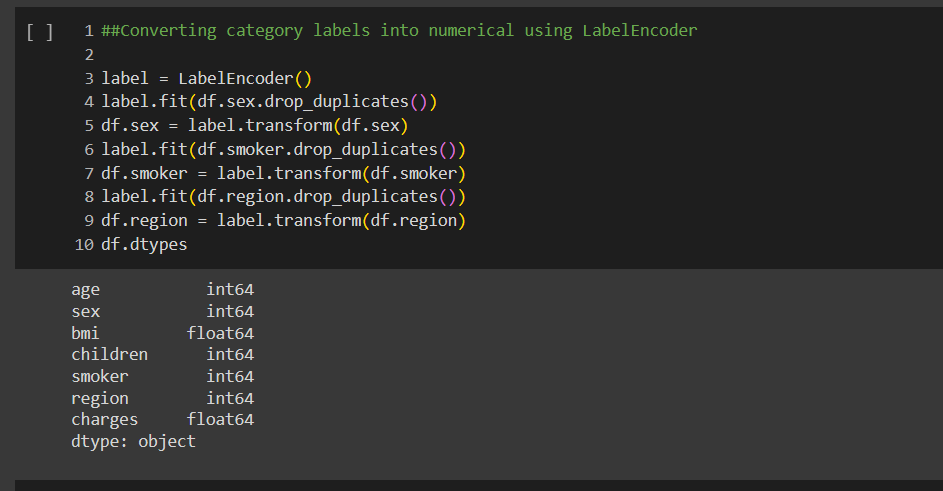
Description automatically generated

The few data processing that was done includes:

1. Conversion of object type data into categorical data types so that they can be encoded.



1. Converting categorical features into numerical representations using labelencoding



1. Model Training

- Defined a list of regression models, including RandomForestRegressor, LinearRegression, Ridge Regression, Lasso Regression, and GradientBoostingRegressor. These models represent a diverse set of regression techniques, each with its own strengths and assumptions.

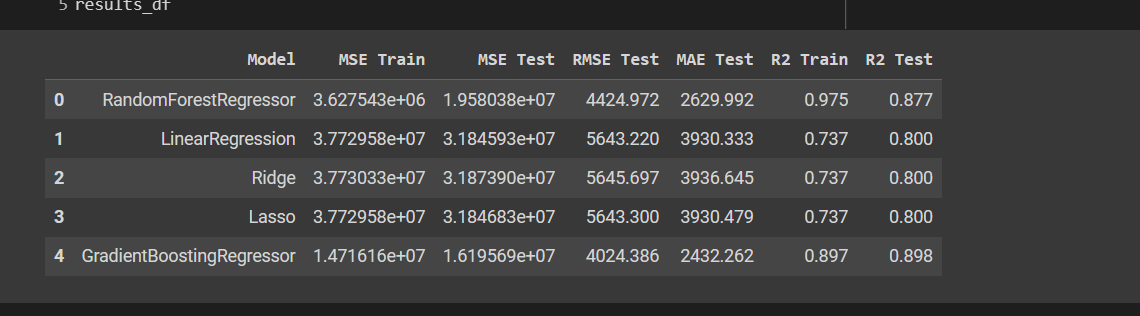
For each regression model, performed the following steps:

1. \*\*Model Training\*\*: Fitted the model on the training data (x\_train, y\_train) to learn the underlying patterns and relationships between input features and target variable (medical costs).

2. \*\*Prediction\*\*: Made predictions on both the training and testing datasets using the trained model. These predictions enabled one to assess the model's ability to generalize to new, unseen data.

3. \*\*Evaluation Metrics\*\*: Computed several evaluation metrics to quantify the performance of each model. These metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score. These metrics provide insights into the accuracy, precision, and goodness of fit of the regression models.

4. \*\*Results\*\*: Stored the evaluation results for each model, including the model name, MSE on training and testing datasets, RMSE on testing dataset, MAE on testing dataset, R2 score on training and testing datasets.



1. Model Selection and Training

In the model selection phase, chose the GradientBoostingRegressor model for its robustness and ability to handle complex relationships within the data and also because it has a higher r2-score on the test set with a value of 0.89 .

\*\*Model Training:\*\*

- Initialized the GradientBoostingRegressor model and fit it on the training data (x\_train, y\_train). During the training process, the model learns the underlying patterns and relationships present in the training dataset.

\*\*Prediction:\*\*

- After training the model, i used it to make predictions on both the training and testing datasets.

- The model predicts medical costs for each instance in the training and testing datasets, providing estimates of healthcare expenses based on individual characteristics and other relevant factors.

\*\*Evaluation:\*\*

- While the specific evaluation metrics are not provided in this snippet, we typically assess the performance of the GradientBoostingRegressor model using standard regression evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score.

- These evaluation metrics helped quantify the accuracy, precision, and goodness of fit of the model, providing insights into its predictive performance on both the training and testing datasets.

1. Model Deployment

- The Flask application initializes an instance of the Flask web framework and loads the pre-trained GradientBoostingRegressor model and label encoder.

- Upon accessing the home route ('/'), the application renders the index.html template, which contains a form for users to input their information.

- When the user submits the form, the application receives a POST request to the '/predict' route.

- In the '/predict' route, the application extracts the input features from the form data, including age, sex, BMI, number of children, smoking status, and region.

- The application performs label encoding on categorical features (sex, smoker, region) using the pre-loaded label encoder.

- Next, the application constructs a feature vector containing the encoded features and reshapes it to match the input shape expected by the model (1 sample, multiple features).

- The feature vector is then passed to the pre-trained GradientBoostingRegressor model for prediction.

- Finally, the predicted medical cost is returned to the user via the 'result.html' template, where it is displayed.

Results and Discussion

Upon evaluating the performance of the GradientBoostingRegressor model on the test dataset, the following metrics were obtained:

- \*\*Mean Squared Error (MSE)\*\*: The model achieved an MSE of approximately 1.4716e+07 on the training dataset and 1.6196e+07 on the test dataset. The MSE measures the average squared difference between the actual and predicted values, with lower values indicating better model performance.

- \*\*Root Mean Squared Error (RMSE)\*\*: The RMSE for the test dataset was approximately 4024.386. RMSE is the square root of the MSE and provides a measure of the average deviation between the predicted and actual values. Lower RMSE values signify better predictive accuracy.

- \*\*Mean Absolute Error (MAE)\*\*: The MAE on the test dataset was approximately 2432.262. MAE represents the average absolute difference between the predicted and actual values, providing insights into the model's prediction accuracy.

- \*\*R-squared (R2) Score\*\*: The R2 score, also known as the coefficient of determination, was approximately 0.897 on the training dataset and 0.898 on the test dataset. The R2 score measures the proportion of the variance in the target variable that is explained by the model. Higher R2 scores indicate better model fit to the data.

\*\*Discussion:\*\*

- The GradientBoostingRegressor model demonstrated strong predictive performance on both the training and test datasets, as evidenced by the low MSE, RMSE, and MAE values.

- The model's high R2 score on the test dataset (0.898) indicates that approximately 89.8% of the variance in medical costs is explained by the model, suggesting a good fit to the observed data.

Conclusion

In conclusion, the deployment and evaluation of the GradientBoostingRegressor model for medical cost prediction have yielded promising results. The model demonstrates strong predictive performance, as evidenced by low mean squared error, root mean squared error, and mean absolute error values on both the training and test datasets. Additionally, the high R-squared score indicates that a significant portion of the variance in medical costs is explained by the model, highlighting its effectiveness in capturing the underlying relationships between input features and healthcare expenses.

The successful deployment of the predictive model offers valuable insights for various stakeholders in the healthcare industry, including policymakers, insurers, and healthcare providers. By accurately estimating medical costs, this model can support informed decision-making, resource allocation, and risk management strategies. Moreover, it provides a valuable tool for individuals to better understand and plan for their healthcare expenses.