MEDICAL INSURANCE COST PREDICTION USING MACHINE LEARNING

Submitted By,

Jerrina Ann Simon

Data Science Intern-Batch B1

Student ID: STB03-T0015

Scifor Technologies

DECLARATION

I hereby declare that the project report entitled '**Medical Insurance Cost Prediction Using Machine Learning'** is a record of my original work It is submitted as part of the training in scifor technologies. The information presented in this project has not been submitted to be assessed for a degree or diploma by any other university or institute.

Name: Jerrina Ann Simon

Student Id: STB03-T0015

Signature: Jerrina

ACKNOWLEDGMENT

I would like to begin by extending my deepest gratitude to my Trainer Urooj Khan for her continuous support and direction throughout the work that would lead me in the right direction. Her contributions and assistance are really important for my project.

Abstract

The healthcare industry, which is responsible for over 30% of the GDP in industrialized nations, has a substantial influence on national economies. This significant expense, both in absolute terms and as a percentage of the overall economy, highlights the crucial role of healthcare expenditures. In particular, through programs like Medicare, which significantly covers medical expenses for older individuals, the government plays a critical role in managing these costs. However, there is an impending economic crisis due to the new generation's impending retirement and subsequently Medicare eligibility. Healthcare prices are increasing, which puts a heavy demand on the government and government resources. To overcome these obstacles, strategic planning is required to keep healthcare costs under control and ensure accessibility and affordability for the growing number of people in need of health services

In this project, we explore the application of machine learning (ML) to enhance the efficiency of insurance policy terms in the insurance industry. The work aims to predict insurance amounts for various types of individuals by combining individual and health data. The project analyses three regression models: linear regression, random forest regression, and gradient boosting regression. These models are trained on a dataset and then predicted using the training data. Our results show that the Gradient Boosting Regression model outperforms the others, with an MAE of 2655.357, MSE of 22457742.416, RMSE of 4738.96, and an impressive R squared value of 0.869. Furthermore, Gradient Boosting and Random Forest emerge as the best-performing models, with R-squared values of 0.869 and 0.863 respectively. This work offers insights on the best methods for showing insurance amounts And deployed it on the streamlit using the best-performing model to predict the health insurance cost based on some factors.

Contents

CHAPTER 1: Introduction

The purpose of this study focuses on informing people about the costs associated with health insurance according to their particular health condition. The aim is to assist people in making well-informed decisions that give priority to important health-related factors above excess costs. The importance of health insurance in today's world cannot be highlighted, since people frequently must deal with the complicated factors affecting insurance premiums, which fluctuate greatly between insurance providers.

In addition, there is a lack of awareness in many rural regions about the free health insurance that the Indian government offers to individuals who fall below the poverty level. Because of this complexity, people living in these locations frequently choose to have private health insurance or decide not to acquire coverage at all, which might leave them vulnerable to health problems. Since our analysis only offers an approximation, it is crucial to emphasize that this amount is not customized for any particular health insurance company. As a general guide, it encourages people to approach their health insurance requirements thoughtfully and ensures they make accurate choices when selecting the right plan.

Problem Statement

There is a serious risk to personal health from the fast-paced and widely accepted behaviours of the modern lifestyle, which highlights the need for innovative ways to deal with the rising cost of healthcare. The fast-paced nature of modern living frequently results in unhealthy behaviours, even with expenditures in preventive health measures. A proactive approach is required to address this problem, specifically the development of an application that informs users about the factors that contribute to health decline and provides an estimate of possible medical costs. With the help of such an application, users would be better equipped to deal with the challenges of rising healthcare expenditures in the fluctuating modern world, promoting financial responsibility and health awareness.

Objective

The main goal of this project is to create a strong machine learning model that can forecast patients' future medical costs. The model attempts to give precise estimates of future medical expenditures by using appropriate factors, enabling active resource allocation and financial planning. The secondary goal is to determine the key factors that influence the predicted medical expenditures to provide an understanding of the elements that determine healthcare

prices. This research aims to improve our understanding of the intricate connection between several factors and medical expenditures by utilizing modern machine-learning techniques. This will eventually help with making informed decisions in the field of healthcare planning and administration.

Chapter 2: Literature Review

This section highlights the current research efforts in the fields of information exploration and machine learning techniques with a particular focus on claim prediction. Jessica Pesantez-Narvaez's 2019 paper, "Utilising Telematics Data to Forecast Automobile Insurance Claims," is a significant addition to this topic. The study compares the effectiveness of XG Boost and logistic regression techniques in forecasting the occurrence of accidents. The results show that logistic regression is more effective than XG Boost because it is easier to understand and predict.

The current analysis uses deep neural networks, machine learning, and advanced statistical approaches to estimate healthcare expenditures, in contrast with earlier research that detected claims-related concerns without considering predicted expenses and claim scope. This is an important step in the direction of using more advanced approaches to improve predictive modeling's knowledge and forecasting of different aspects related to healthcare and insurance. By offering deeper insights and more precise forecasts in the challenging field of healthcare cost prediction, the combination of these cutting-edge methodologies aims to contribute an important contribution to the predictive modelling environment as it keeps on changing.

CHAPTER 3: Methodology

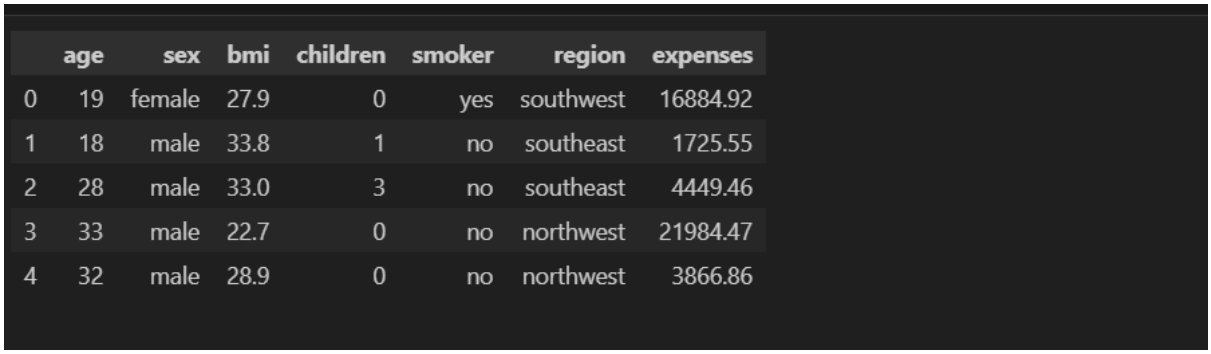
Dataset

This project uses data from Kaggle, which is a dataset consisting of 1338 rows and 7 columns. The dataset includes essential data on medical expenses, with the columns representing age, sex, bmi, children, smoker, region, and expenses. 'Expenses' is the target column as the primary objective is to predict medical costs; the other columns are independent variables that feed into the prediction model.

The influence of the 'age' column on health outcomes makes it an important component to consider. As an average, younger people often have fewer medical costs than elderly people. The 'sex' column notes that men and women have different healthcare demands and presents gender as a possible determining factor. Subsequently, the 'BMI' column indicates the Body Mass Index, an important health indicator. Extreme BMI holders those who are too high or too

low are more likely to have more medical expenses as they usually need more maintenance. The 'children' column offers information on how many kids an individual has, recognizing the extra cost associated with childcare responsibilities and how it affects overall medical costs.

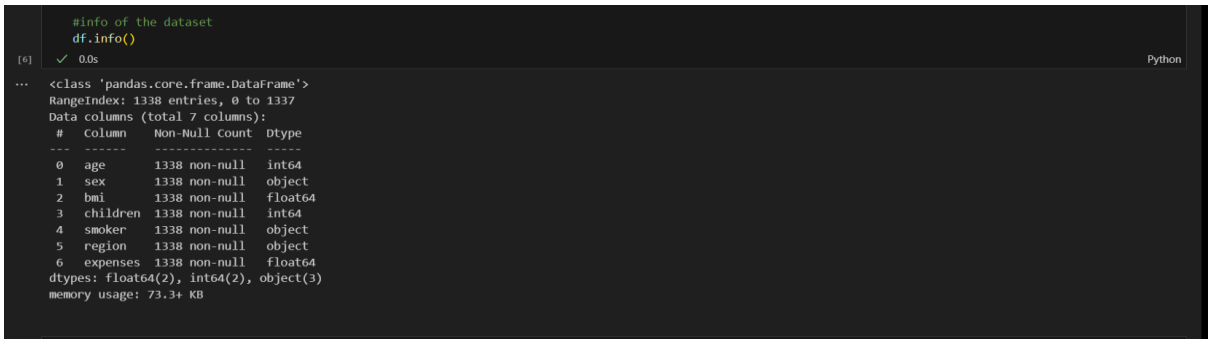
The 'smoker' column highlights how important smoking habits are for forecasting medical costs. Smoking is a significant risk factor, particularly as people become older, thus it is an important component of our prediction model. The 'region' column, which acknowledges the influence of environmental variables on health, brings geographical issues forward. Wealthy and hygienic areas might lead to cheaper medical expenditures, whereas unfavourable circumstances can lead to higher healthcare costs. This diverse dataset provides the groundwork for developing a strong prediction model to improve our understanding of the complicated factors affecting medical costs.



	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86

Data Preperation

The initial exploration of the provided DataFrame (**df**) involved several key operations to understand its characteristics. The dimensions of the dataset were unveiled using the **df.shape** command, indicating 1338 rows and 7 columns. Subsequently, **df.info()** provided concise information about data types and non-null counts for each column.



```
#info of the dataset
df.info()
```

[6] ✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   expenses  1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

while **df.dtypes** gives the details of the specific data types allocated to each. A comprehensive statistical overview, encompassing both numerical and categorical columns, was generated using **df.describe(include="all")**.

```
#summary of the dataset
df.describe(include="all")
```

[8] ✓ 0.0s Python

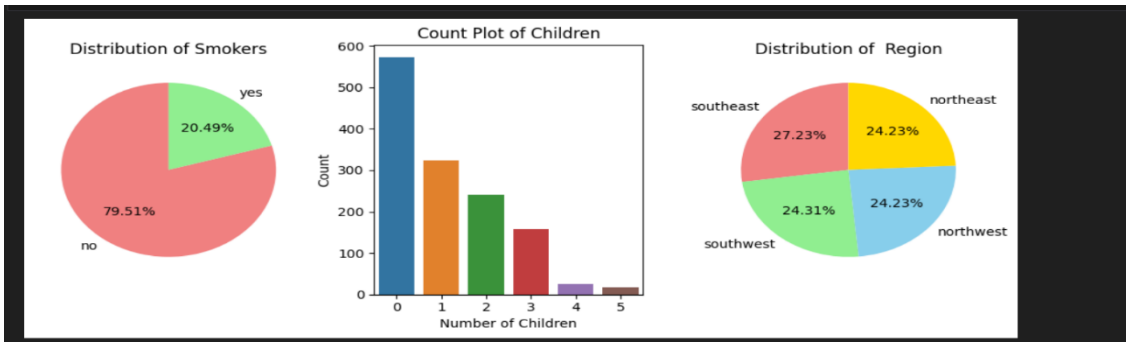
	age	sex	bmi	children	smoker	region	expenses
count	1338.000000	1338	1338.000000	1338.000000	1338	1338	1338.000000
unique	NaN	2	NaN	NaN	2	4	NaN
top	NaN	male	NaN	NaN	no	southeast	NaN
freq	NaN	676	NaN	NaN	1064	364	NaN
mean	39.207025	NaN	30.665471	1.094918	NaN	NaN	13270.422414
std	14.049960	NaN	6.098382	1.205493	NaN	NaN	12110.011240
min	18.000000	NaN	16.000000	0.000000	NaN	NaN	1121.870000
25%	27.000000	NaN	26.300000	0.000000	NaN	NaN	4740.287500
50%	39.000000	NaN	30.400000	1.000000	NaN	NaN	9382.030000
75%	51.000000	NaN	34.700000	2.000000	NaN	NaN	16639.915000
max	64.000000	NaN	53.100000	5.000000	NaN	NaN	63770.430000

Potential missing values were examined using **df.isnull().sum()** in order to improve data quality. To further improve the accuracy of the data, the dataset was improved using **df.drop_duplicates(),** which eliminated duplicate rows. The modified dimensions were confirmed by the final verify, **df.shape** post-duplicate elimination, which resulted in a thorough review necessary for further analytical efforts. This methodical technique guarantees that the dataset is understood and ready for more in-depth examination in later project stages.

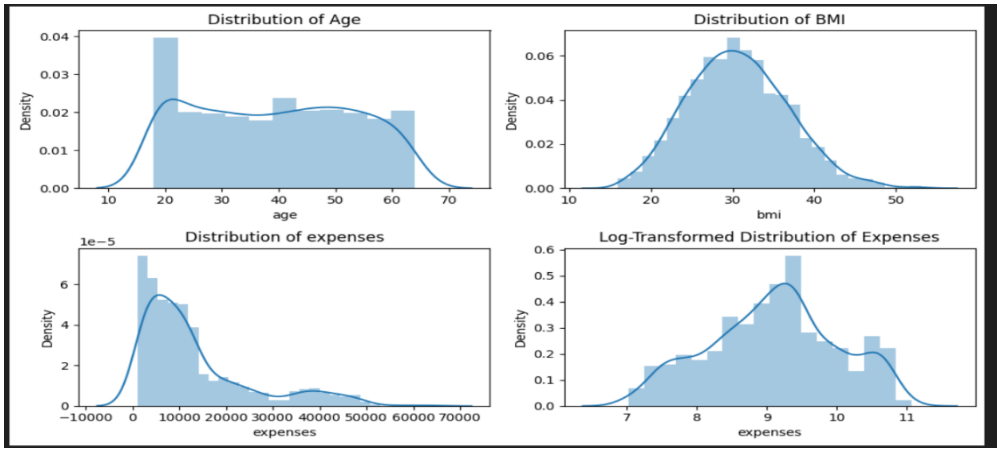
Data Visualization

During the univariate analysis stage, analysed certain dataset columns using numerical summaries and visualizations. About the 'Smoker' column, a pie chart explained that 20.49% of participants smoke, while 79.51% of participants do not smoke. As one moved to the 'Children' column, the distribution of subjects with different numbers of children (from 0 to 5) was shown graphically by a Count plot. Upon study, it was found that the largest number of subjects 574 were childless, with 324 having one, 240 having two, 157 having three, 25 having four, and 18 having five children. A pie chart that followed offered a more detailed look at the distribution of the regions, showing that the Northwest and Northeast had the same percentage of 24.23%. More in-depth analysis in later parts of the report will be developed on the basis this univariate examination offers regarding how it is organized and distributed within

those columns.

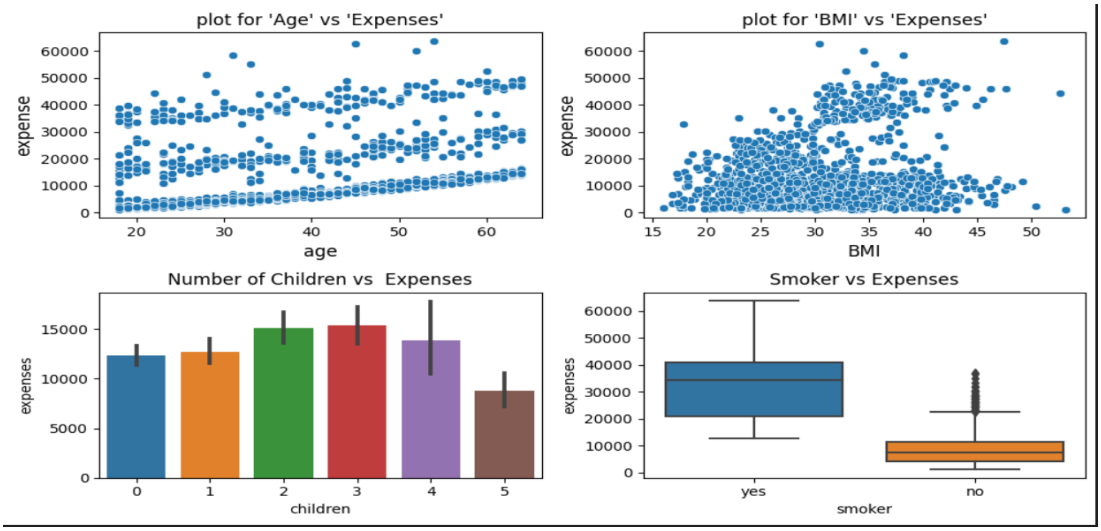


With distribution plots, the analysis of distribution focused on "Age," "BMI," and "Expenses," showing meaningful patterns. All age groups had equally distributed representation, according to the 'Age' distribution. A mostly normal distribution with outliers at both lower and upper ends was observed for "BMI," with the central point of the distribution around 30. On the other hand, the 'Expenses' column showed a pattern that was skewed to the right, which led the use of a log transformation for normalization by modeling assumptions. These results give a concise summary of the variable distributions in the dataset, directing next analytical procedures and ensuring an in-depth understanding of the properties of the data for effective predictive modelling.

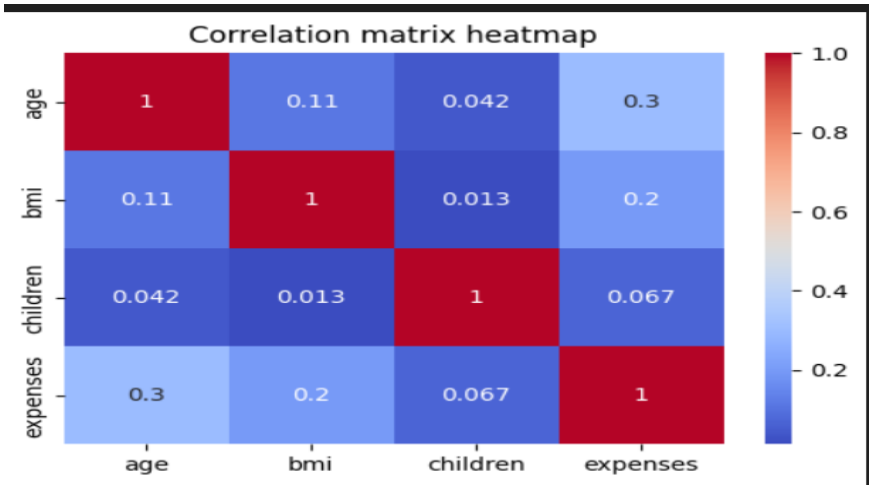


In the bivariate analysis, scatter plots were used to investigate correlations between key variables. The 'Age' versus 'Expenses' plot showed that medical costs generally rise as people age, yet some elderly people had lower costs than others. scatterplot application made it possible to create lines of trend which show a linear connection. A similar method was used to compare "BMI" to "Expenses," showing an uneven pattern but indicating that those with higher BMIs will have greater medical costs.

Bar charts showed that the number of "Children" and "Expenses" were positively correlated, with people who had three children spending the most money. Box plots showed that compared to non-smokers, smokers often had far greater medical costs. These revelations improve our knowledge of variable relationships, which is important for future predictive modelling and decision-making procedures.



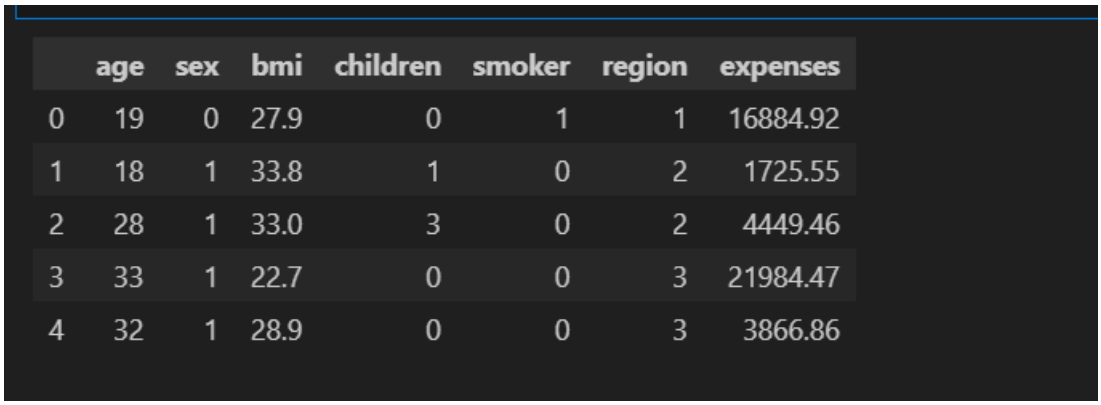
The correlation matrix's produced heatmap offers a graphic depiction of the connections between the dataset's numerical variables. The correlation coefficient between two variables is represented by each cell in the heatmap, and the direction and intensity of these associations are shown by colour gradients. Warmer colours are used to illustrate positive correlations, whereas cooler colours are used to illustrate negative correlations. Accurate correlation values are provided by the numerical values that are included in each cell. In data exploration and machine learning operations, this visualisation is especially helpful for seeing possible patterns, dependencies, or multicollinearity between variables. It also helps with feature selection, model development, and subsequent analyses.



Data Preprocessing

Encoding of categorical column

The dataset includes three categorical columns that convert the categorical variables in the Data Frame (df) into numerical representations: "sex," "smoker," and "region." In the 'sex' column, its maps' male' to 1 and 'female' to 0, translating gender into numerical numbers. Similar to this, the 'smoker' column maps 'yes' to 1 and 'no' to 0, making it easier to describe smoking status numerically. To translate geographic data into a number format, the 'region' column is mapped, with 1 denoting the southwest, 2 the southeast, 3 the north west, and 4 the northeast. By adding categorical data into machine learning models, this kind of encoding improves the readability and effectiveness of later analyses.



	age	sex	bmi	children	smoker	region	expenses
0	19	0	27.9	0	1	1	16884.92
1	18	1	33.8	1	0	2	1725.55
2	28	1	33.0	3	0	2	4449.46
3	33	1	22.7	0	0	3	21984.47
4	32	1	28.9	0	0	3	3866.86

Train Test split

separates the dataset into the target variable (y) and independent variables (x). Following that, it splits the data into training and testing sets, using 20% of the data for testing (x_test and y_test) and 80% of the data for training (x_train and y_train). This division is crucial for training and evaluating machine learning models. The random seed (random_state=42) ensures consistent results in the split.

Feature Scaling

feature scaling standardizes the range of independent variables in a dataset. Using training data, the fit_transform function is used to identify patterns and scale features. The transform function is used for testing data in order to maintain consistency without adding information from the testing set to the training set.

Hypothesis Testing

t-test

Potential relationships were explored by conducting t-tests on the 'sex' and 'smoker' columns with respect to the target variable 'expense'. The Null Hypothesis predicted that the mean

expenses for both genders would be identical for the 'sex' column (male vs. female). However, the Null Hypothesis was rejected due to the derived p-value of 0.033, which was below the 0.05 significance level and indicated a substantial difference in mean expenses between males and females. Similarly, the Null Hypothesis assumed equal mean expenditures for the 'smoker' column (smoker vs. nonsmoker). The Null Hypothesis was rejected due to the very low p-value (1.406724235056543e-282), which is significantly less than 0.05 and shows a significant difference in mean expenses between smokers and nonsmokers. These results show the significance it is for "sex" and "smoking" status to have an effect on health care expenses.

ANOVA Test

The ANOVA test is used to evaluate if the means of three or more groups differ significantly based on statistical significance, providing useful data about how different category factors might affect the numerical outcome. 'children' and 'region' columns were subjected to ANOVA testing using the 'expense' target variable. Regarding 'children,' the rejection of the Null Hypothesis (p-value = 0.00378) showed that the mean expenses changed amongst those who had no children, one children, two children, and three children. The Null Hypothesis for "region," which was rejected (p-value = 0.0327), indicated variations in mean expenses between the southeast, northeast, northwest, and southwest areas. These findings highlight how "children" and "region" affect health care costs.

Chi-Square Test

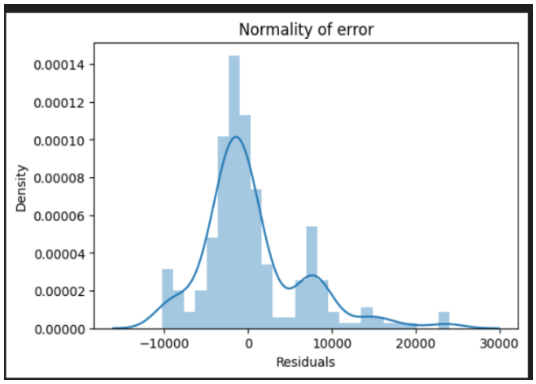
To determine if there is any correlation between categorical variables, the chi-square test is performed. To evaluate any correlations, the chi-square test was used to the 'Sex' and 'Smoker' columns. There was no association between these category variables, according to the null hypothesis. But the Null Hypothesis was rejected with a p-value of 0.006 (below 0.05) and a chi-square statistic value of 7.46 (beyond the threshold limit of 3.84). This highlights the connection between these categorical variables by indicating a statistically significant association between the 'Sex' and 'Smoker' columns.

Model Building

Linear Regression

A statistical method used to represent the connection between independent and dependent variables is called linear regression. Here uses scikit-learn to show linear regression and evaluates the model's performance using metrics like mean absolute error, mean squared error, root mean square error, and R-squared score. It uses stats models to provide a complete

statistical summary, generates a DataFrame to compare actual and predicted values, and evaluates the residuals' normality using a histogram. A graph is plot for showing the first 250 actual and predicted values. This deep dive offers an in-depth understanding of the linear regression process by that covers model training, evaluation, statistical summary and visualisation.



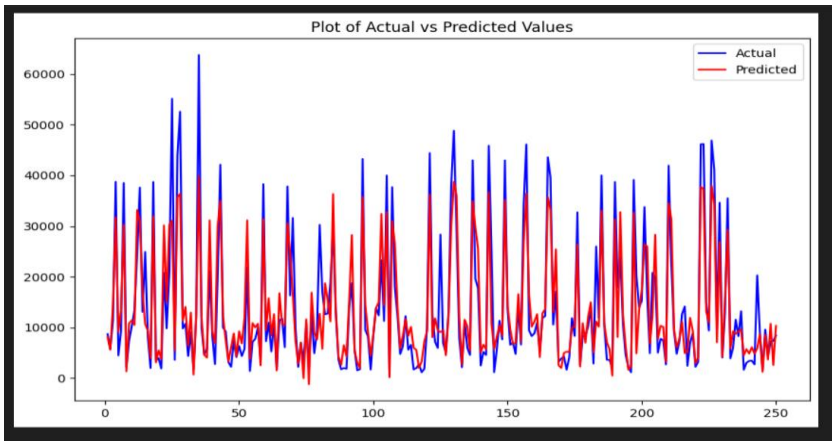
OLS Regression Results

Dep. Variable:	expenses	R-squared:	0.730
Model:	OLS	Adj. R-squared:	0.728
Method:	Least Squares	F-statistic:	477.9
Date:	Mon, 05 Feb 2024	Prob (F-statistic):	1.41e-297
Time:	14:41:10	Log-Likelihood:	-10831.
No. Observations:	1069	AIC:	2.168e+04
Df Residuals:	1062	BIC:	2.171e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.303e+04	186.659	69.808	0.000	1.27e+04	1.34e+04
x1	3480.3270	188.643	18.449	0.000	3110.172	3850.482
x2	-49.9280	187.659	-0.266	0.790	-418.153	318.297
x3	1892.0041	190.770	9.918	0.000	1517.676	2266.333
x4	638.2386	187.033	3.412	0.001	271.243	1005.235
x5	9223.7424	187.514	49.190	0.000	8855.802	9591.683
x6	262.2355	189.175	1.386	0.166	-108.965	633.436

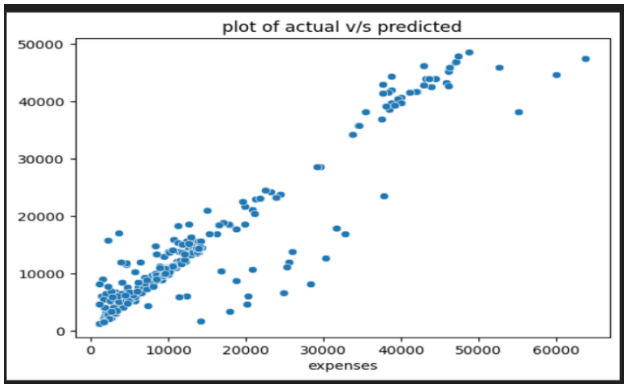
Omnibus:	264.931	Durbin-Watson:	1.966
Prob(Omnibus):	0.000	Jarque-Bera (JB):	635.222
Skew:	1.324	Prob(JB):	1.16e-138

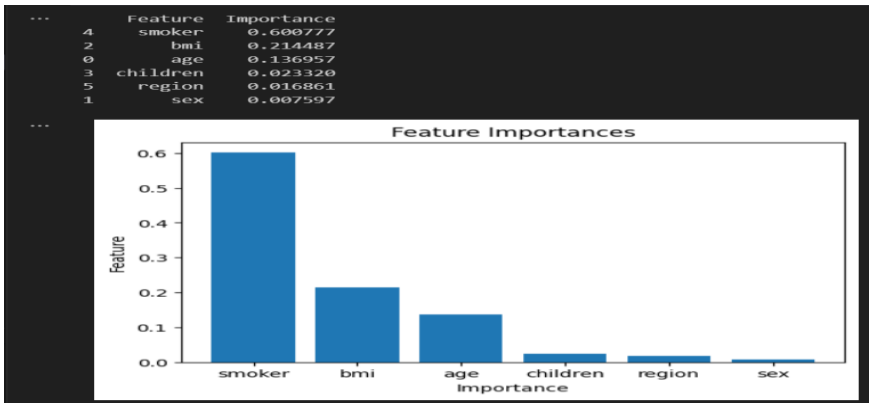
...



Random Forest Regressor

The Random Forest Regressor is an ensemble learning approach, combining multiple decision trees to enhance predictive accuracy. uses scikit-learn to create a Random Forest Regressor. The model is trained on the provided data, and its performance is assessed using metrics such as mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R-squared (R2) score. Using a scatter plot, further illustrates the expected values in comparison to the actual values. It also computes and shows the feature importances, giving information on how each feature affects the model's predictions. The Random Forest Regressor is a reliable method that can handle complex relationships in data, which makes it suitable for various kinds of regression tasks.

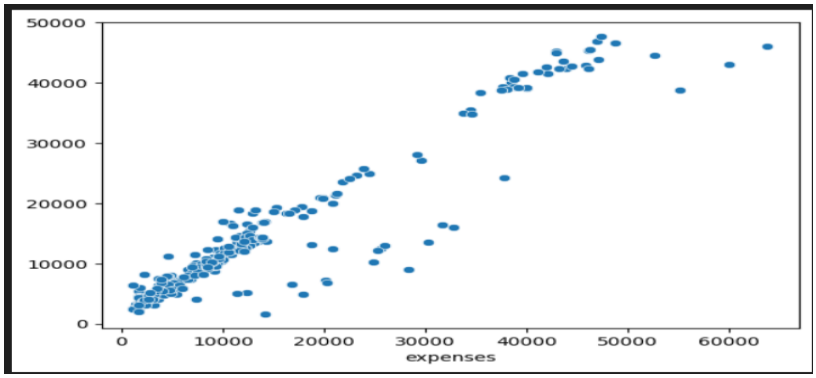


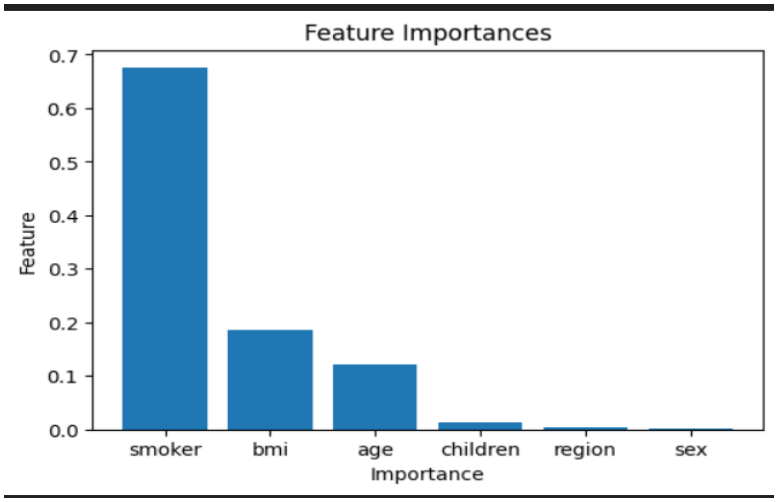


Gradient Boosting Regressor

A Gradient Boosting Regressor using scikit-learn, a powerful ensemble learning algorithm that builds decision trees sequentially, with each tree correcting the errors of the previous one. The model is trained on the given data and evaluated using metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score. Gradient Boosting Regressor is known for its ability to capture complex relationships in data and provide accurate predictions. Provides a scatter plot visualizing the predicted values against the actual values. Additionally, the feature importances are calculated and displayed,

offering insights into the contribution of each feature to the model's predictions. Gradient Boosting Regressor is a versatile algorithm suitable for various regression tasks, providing robust and accurate results.





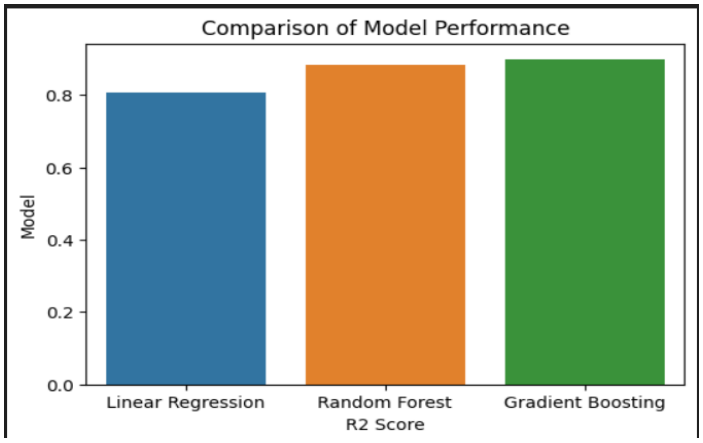
Cross-Validation

When assessing machine learning models on small data sets, cross-validation is a crucial resampling method. Using this approach, models are iteratively trained and tested with 'k' set to 5. Three models were evaluated using cross-validation in this project and the Gradient Boosting Regressor had the highest mean score of 0.833. This result indicates that, out of all the models examined, the Gradient Boosting Regressor is the most successful in predicting medical expenses. For accurate model evaluation and accurate model selection, cross-validation is necessary.

Comparing Three models

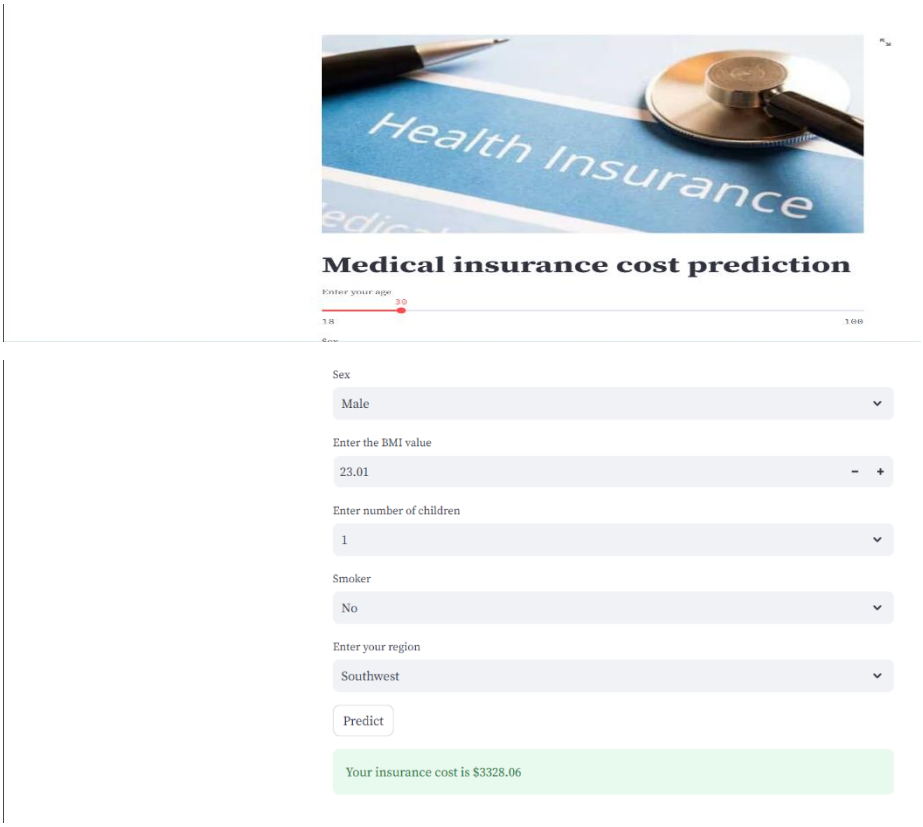
Creates arrays for R2 scores and corresponding labels for three different regression models: Linear Regression, Random Forest, and Gradient Boosting. The bar plot of R2 scores reveals that the Gradient Boosting Model excels among the three models, making it the best choice for predicting medical expenses. Similarly, the Linear Regression Model performs the least

effectively. This comparison underscores the successful development and evaluation of predictive models, giving the importance of selecting the most accurate algorithm for medical expense prediction.



Deployment

Streamlit framework is used for deploying a medical insurance cost prediction model. The trained model, stored as '**trained_model.joblib_1'**, is loaded into the application. The user interacts with a user-friendly interface on the homepage, where they input information such as age, sex, BMI, number of children, smoking status, and region. Upon clicking the 'Predict' button, the model processes the input data and provides a real-time prediction of the insurance cost. The app has been implemented through streamlit therefore the users can get the medical expense by accessing through the link of the app.



Chapter 4: Conclusion

The project addresses the prediction of medical expenses based on factors such as age, gender, BMI, smoking habits, number of children, and region. It explores the dataset through visualizations and statistical tests, highlighting the significance of smoking behaviour and age in influencing healthcare costs. Three models (Linear Regression, Random Forest, Gradient Boosting) are developed and compared, with Gradient Boosting Regressor proving to be the most effective. The models are evaluated using metrics like R-squared and RMSE. The deployment done using the streamlit allows users to predict medical expenses The analysis and models developed provide insights into the complex relationship between various factors and medical expenses, contributing to a better understanding of healthcare cost prediction.

Reference

[1] https://www.kaggle.com/code/abdulrahmanelbanna/medical-insurance-cost-prediction/input

[2] https://www.youtube.com/watch?v=ntBa7YKc9XM

[3] https://www.geeksforgeeks.org/medical-insurance-price-prediction-using-machine-learning-python/

[4] https://www.analyticsvidhya.com/blog/2021/05/prediction-of-health-expense/

[5]https://www.researchgate.net/publication/374553777_Medical_Insurance_Cost_Prediction_Using_Machine_Learning