Team – 7
DATS6103: Summary Report
Prof. Dr. Ning Rui
Date: 12/14/2023

PREDICTING HEALTH INSURANCE PREMIUMS IN THE U.S.

I. Introduction

Insurance is a safety net for people and organizations in the event of unforeseen calamities or disasters, making it an essential component of contemporary financial planning. Health insurance is a type of coverage that helps individuals manage and mitigate the financial risks associated with medical expenses. It functions as a legal contract between an insurance provider and a policyholder, whereby the provider agrees to pay regular premiums in return for the insurer agreeing to pay a percentage of the policyholder's medical expenses. A variety of medical services, including doctor visits, hospital stays, surgeries, prescription drugs, preventive care, and other medically required treatments, can be covered by health insurance.

Debugging some interesting facts about Health Insurance in the United States. While the majority of Americans have health insurance, a notable portion still grapples with medical debt. In 2022, the U.S. Census Bureau noted an increase in health insurance coverage, with 92.1% of the population having coverage, up from 91.7% in 2021. Despite this, approximately 8.4% or 27.6 million American adults faced periods without healthcare coverage in 2022. A worrisome revelation indicates that nearly 25% of adults acknowledged skipping medication doses, cutting pills, or not filling prescriptions in the past year due to cost concerns. Moreover, about 41% of adults reported having outstanding medical or dental bill debt. Dental services ranked as the most frequently delayed healthcare type due to costs (35%), followed by vision services (25%) and doctor's visits (24%). These statistics underscore persistent challenges in achieving comprehensive and accessible healthcare for all Americans.

Evidently, the pivotal determinant shaping individuals' decisions regarding health insurance is the cost, a variable that is neither rigid nor malleable but contingent upon diverse factors including age, gender, medical history, exercise routines, and smoking habits. The insurance cost for each individual exhibit’s variation based on these nuanced factors.

In our project, a comprehensive analysis of insurance data from a diverse range of individuals was conducted. Leveraging various sophisticated machine learning models, we endeavored to fine-tune and optimize the costs for individuals based on specific and discerning criteria.

We employed diverse visualization techniques to illustrate distinct factors and conducted statistical tests to assess the significance of variables in relation to the cost. Subsequently, we utilized machine learning models for further analysis.

II. SMART Questions

We formulated several SMART goals to guide our research and endeavored to address them through our comprehensive analysis. The following are the questions we established:

→ To what extent is "region" a useful variable for estimating insurance costs? Does the data show any regional trends that affect premiums? Else what specific factors affect premiums?

→ In comparison to non-smokers, how much does being a "smoker" add to the rise in insurance costs?
→ How much does age impact insurance premiums, and is this impact consistent across different regions?
→ How can insurance companies use the data on smoking habits and exercise frequency to devise strategies for premium adjustments?
→ How relevant is gender in determining insurance premiums, and is there a gender-based disparity in premiums?
→ How can we provide individuals with real-time estimates of their health insurance premiums based on their unique characteristics beforehand?

III. Literature Review

As international students embarking on our academic journey in the United States, this project is rooted in our collective experiences upon arrival. The initial phase posed significant challenges, particularly in selecting an appropriate insurance provider. Our primary focus centered on two critical aspects: cost and coverage. As students, our financial resources were inherently constrained, magnifying the importance of securing comprehensive coverage. Given the exorbitant costs associated with emergencies in the United States, the selection of an insurance provider became a crucial and intricate decision-making process. Kaushik et al. [1] have performed research on predicting the insurance cost using an Artificial Neural Network (ANN) -based research regression model.

IV. Description of Data

Our data was obtained from Kaggle and is formatted as a comma-separated values (CSV) file. It comprises 1 million records distributed across 12 columns. Each row in the dataset represents a distinct record, while each column corresponds to a different variable. The variables encompass age, sex, BMI, number of children, smoker status, region, income, education, occupation, and type of insurance plan. Notably, the "charges" column denotes the actual premium amount charged.

#	Column	Dtype
1	age	int64
2	gender	object
3	bmi	float64
4	children	int64
5	smoker	object
6	region	object
7	medical_history	object
8	family_medical_history	object
9	exercise_frequency	object
10	occupation	object
11	coverage_level	object
12	charges	float64

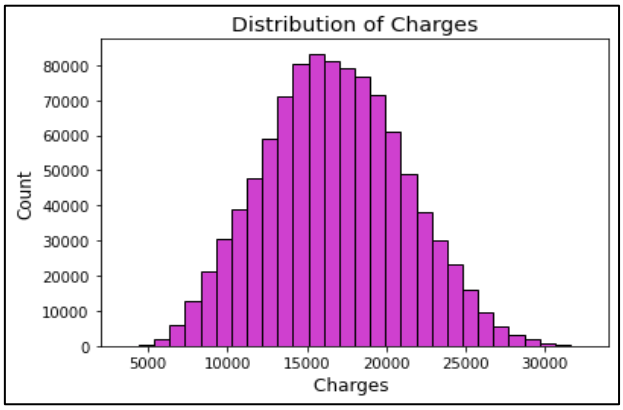
V. Data Preprocessing

Our data preprocessing involved systematically identifying and removing duplicates, addressing missing values with a thoughtful imputation strategy, and transforming categorical

variables through efficient encoding techniques. These streamlined steps laid the groundwork for a refined dataset ready for subsequent analysis and modelling.

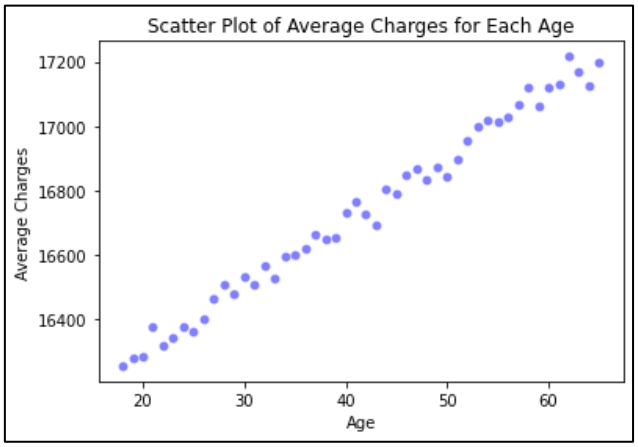
VI. Exploratory Data Analysis

Charges Distribution:



The histogram depicts the charges distribution, showcasing a range spanning from a minimum of 5000 to a maximum of 30000 and exhibiting characteristics reminiscent of a normal distribution. Moving forward, our analysis will focus on understanding how these charges are applied, considering a variety of influencing factors.

Average charges for each age:

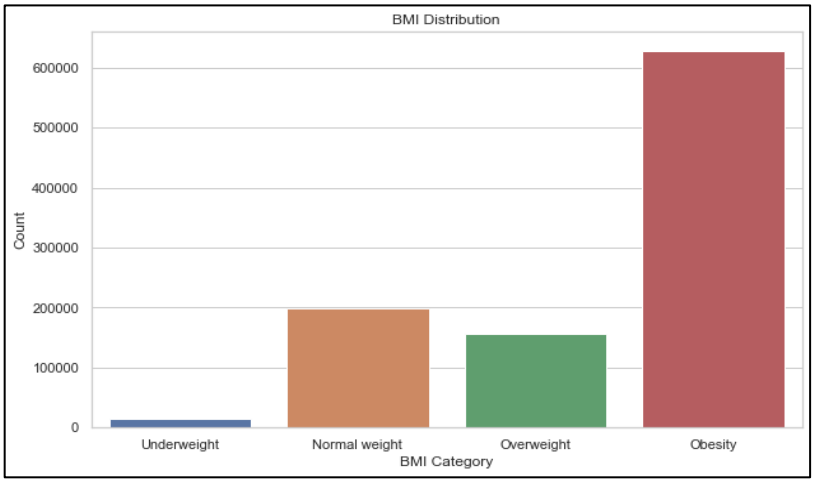


The scatter plot illustrates the correlation between age and mean charges. It indicates that as age rises, so do the charges, resembling a straightforward linear relationship between age and mean charges.

Body Mass Index (BMI):

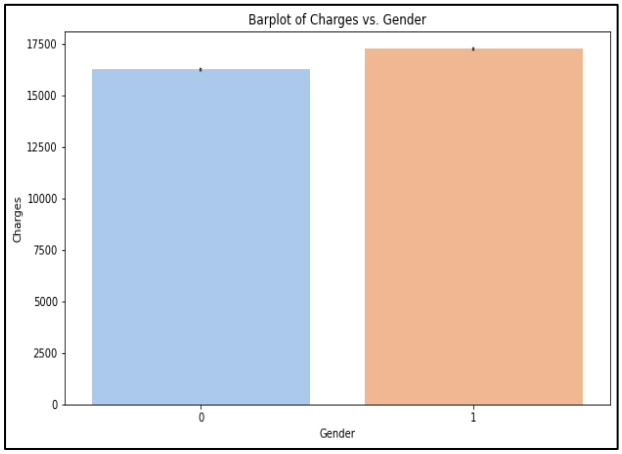
Body Mass Index (BMI) quantifies the amount of body fat relative to an individual's height and weight. BMI is classified on the following basis

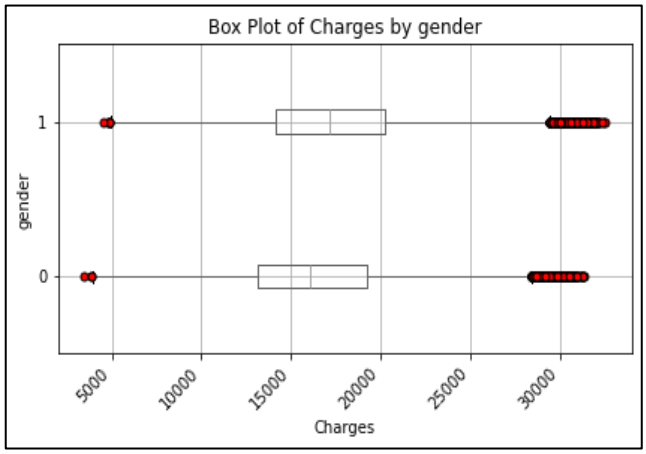
BMI	Category
< 18.5	Under Weight
18.50 - 24.90	Normal Weight
25.00 - 29.90	Overweight
>=30	Obesity



The bar plot above illustrates the distribution of BMI categories among individuals in the dataset. A significant majority are dealing with obesity. In line with 2021 statistics, a noteworthy 40% of Americans are currently grappling with obesity, and this issue is escalating at an alarming rate, posing a concerning trend for humanity.

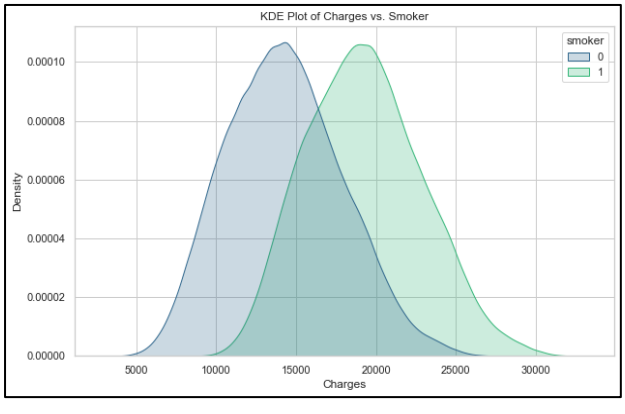
Charges by gender:

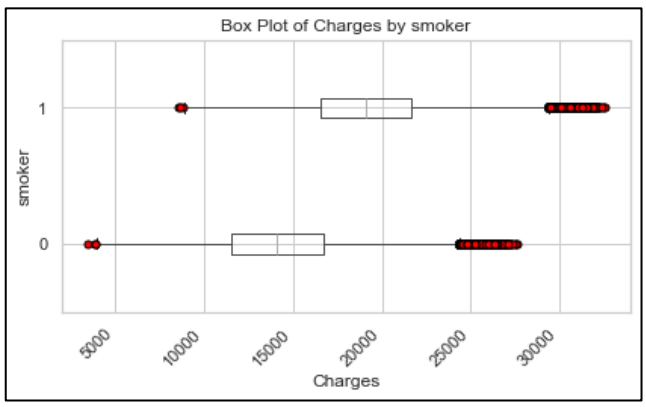




The above plots highlight the pricing contrast between male (1) and female (0). On average, males tend to incur higher charges than females. Notably, the dataset's maximum charge is attributed to a male, while females are associated with the minimum charges. In our investigation, a t-test comparing charges based on gender revealed an extremely low p-value (0.0), providing strong evidence that a significant difference exists in charges between genders. Importantly, our analysis rejects the null hypothesis.

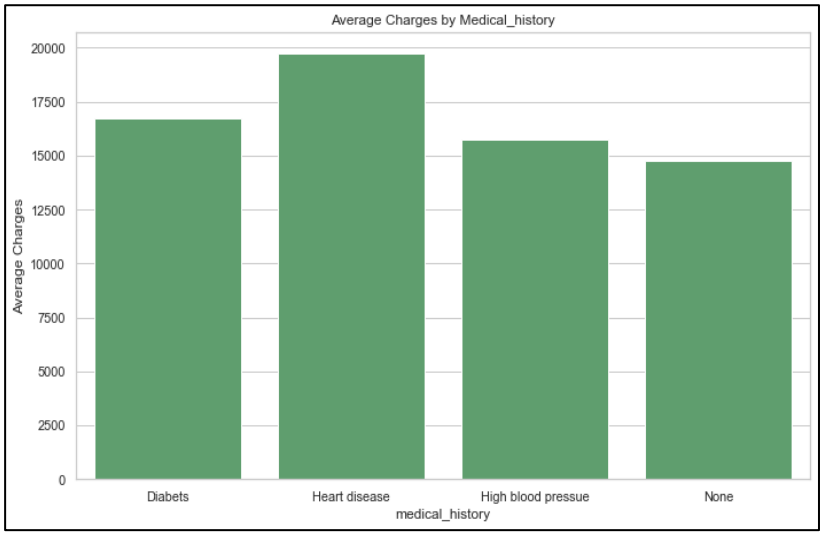
Charges by Smoker:





The KDE and Box plots distinctly show that individuals who smoke generally face higher charges compared to non-smokers. To reinforce this observation, we conducted a T-test, resulting in an exceptionally low p-value ($p < 0.05$). This confirms a significant difference in charges between smokers and non-smokers, and notably, we reject the null hypothesis.

Charges by Medical History:



The plot above delineates the connection between charges and an individual's medical history. It distinctly shows that individuals with a history of "Heart disease" incur notably higher charges, while the distinctions for Diabetes and High Blood Pressure are comparatively marginal.

The groups are labelled as follows:
Diabetes (0), Heart disease (1), High blood pressure (2), and None (3).

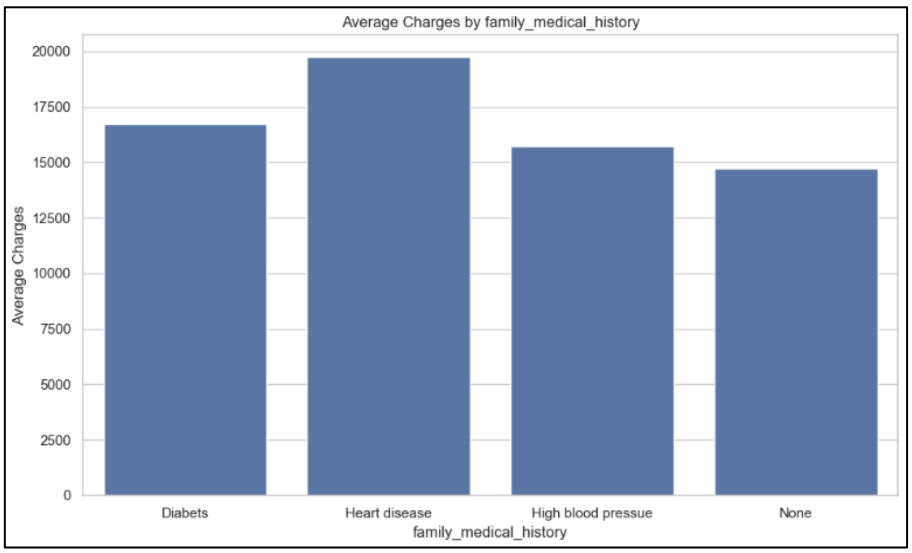
The presented analysis is an ANOVA table, indicating a significant difference among the means of the medical history groups concerning charges. The Tukey HSD post hoc test reveals specific pairwise mean differences and associated confidence intervals. Notably, all group comparisons demonstrate statistically significant differences ($p < 0.05$).

```
                     df      sum_sq     mean_sq       F     PR(>F)
C(medical_history)      3.0  3.515e+12  1.172e+12  73302.91    0.0
Residual           999996.0  1.598e+13  1.598e+07       NaN    NaN

 There is significant difference among the means of medical_history groups of charges
    Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================
group1 group2  meandiff  p-adj   lower      upper    reject
-------------------------------------------------------------
     0      1  2993.0536 0.001  2963.9868  3022.1203   True
     0      2  -1023.212 0.001 -1052.2886  -994.1354   True
     0      3 -2009.9191 0.001 -2038.9673 -1980.8709   True
     1      2 -4016.2656 0.001 -4045.3193 -3987.2118   True
     1      3 -5002.9727 0.001  -5031.998 -4973.9474   True
     2      3  -986.7072 0.001 -1015.7423   -957.672   True
-------------------------------------------------------------
```

Charges by Family Medical History:

The chart below depicts the relationship between charges and an individual's family medical history. Heart disease incurs the highest charges, followed by diabetes and high blood pressure. Conversely, the category "None" signifies the absence of the mentioned diseases, reflecting comparatively lower charges.



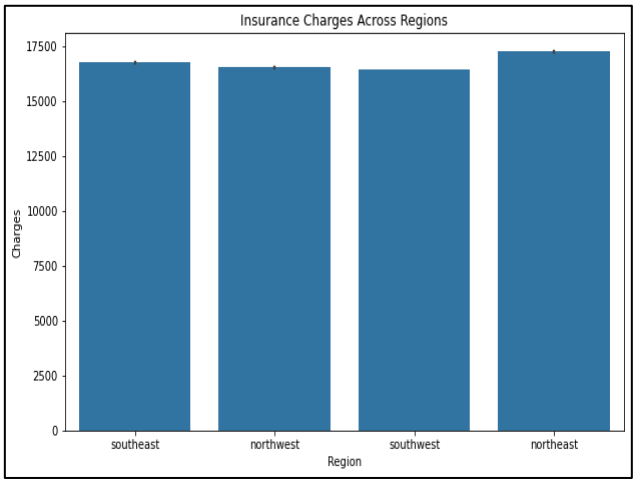
ANOVA results below highlight significant differences in mean charges among family medical history groups (Diabetes, Heart disease, High blood pressure, None). Tukey HSD post hoc tests confirm pairwise distinctions, all with $p < 0.05$, underscoring the significant impact of family medical history on charges.

```
                              df        sum_sq       mean_sq            F  \
C(family_medical_history)      3.0  3.514954e+12  1.171651e+12  73299.432622
Residual                  999996.0  1.598439e+13  1.598445e+07          NaN

                           PR(>F)
C(family_medical_history)     0.0
Residual                      NaN

 There is significant difference among the means of family_medical_history groups of charges
               Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================================================
       group1             group2        meandiff  p-adj   lower      upper    reject
---------------------------------------------------------------------------------------------
          Diabetes       Heart disease  3005.2492 0.001  2976.1915  3034.3068   True
          Diabetes High blood pressure -1005.1294 0.001 -1034.1932  -976.0656   True
          Diabetes               None -2002.7428 0.001 -2031.7898 -1973.6959   True
     Heart disease High blood pressure -4010.3786 0.001 -4039.4337 -3981.3234   True
     Heart disease               None  -5007.992 0.001 -5037.0303 -4978.9537   True
High blood pressure              None  -997.6134 0.001 -1026.6578   -968.569   True
```

Charges across regions:

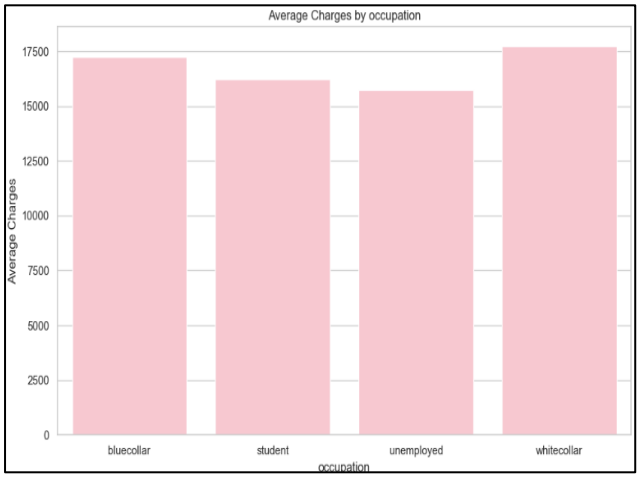


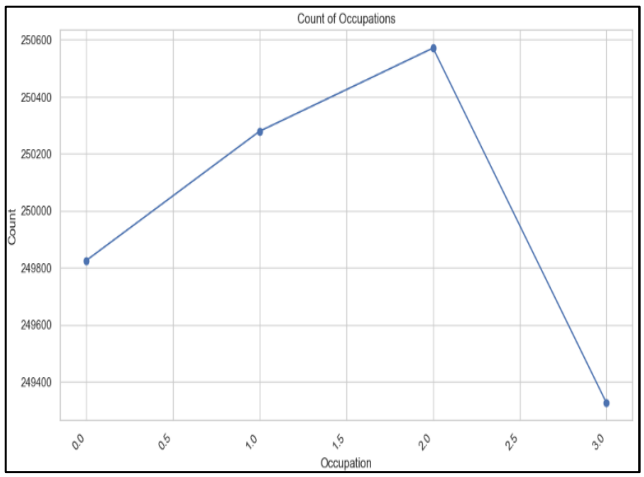

```
F-statistic: 1644.2059645503782
P-value: 0.0
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================
  group1    group2   meandiff p-adj   lower     upper   reject
-------------------------------------------------------------
northeast northwest -707.2576 0.001 -739.2566 -675.2586   True
northeast southeast -497.3613 0.001 -529.3699 -465.3527   True
northeast southwest -799.1973 0.001 -831.1911 -767.2036   True
northwest southeast  209.8963 0.001  177.8745  241.9181   True
northwest southwest  -91.9397 0.001 -123.9466  -59.9328   True
southeast southwest  -301.836 0.001 -333.8526 -269.8195   True
```

The charges plot across regions reveals slight variations, with the North East exhibiting comparatively higher charges. An ANOVA test confirms significantly different mean insurance charges among the four regions, supported by an extremely low p-value. Subsequent Tukey HSD analysis identifies specific pairwise differences in mean charges.

Charges by occupation:

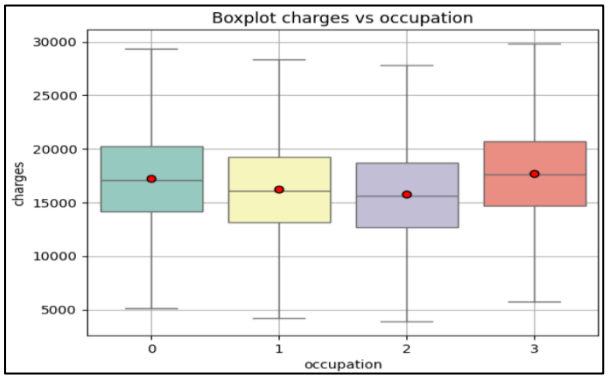
The plots above and box plot below clearly indicate higher charges for individuals in white-collar occupations, followed by those in blue-collar and student roles. Unemployed individuals, while paying less in charges, constitute a higher count in the dataset.





The ANOVA test underscores a substantial difference in mean charges across diverse occupations. Additionally, the Tukey HSD test affirms statistical significance in all pairwise differences. Specifically, the smallest difference is observed between Group 1 (students) and Group 2

(unemployed), with an average charge difference of $495.5484, indicating the varying impact of occupation on charges.



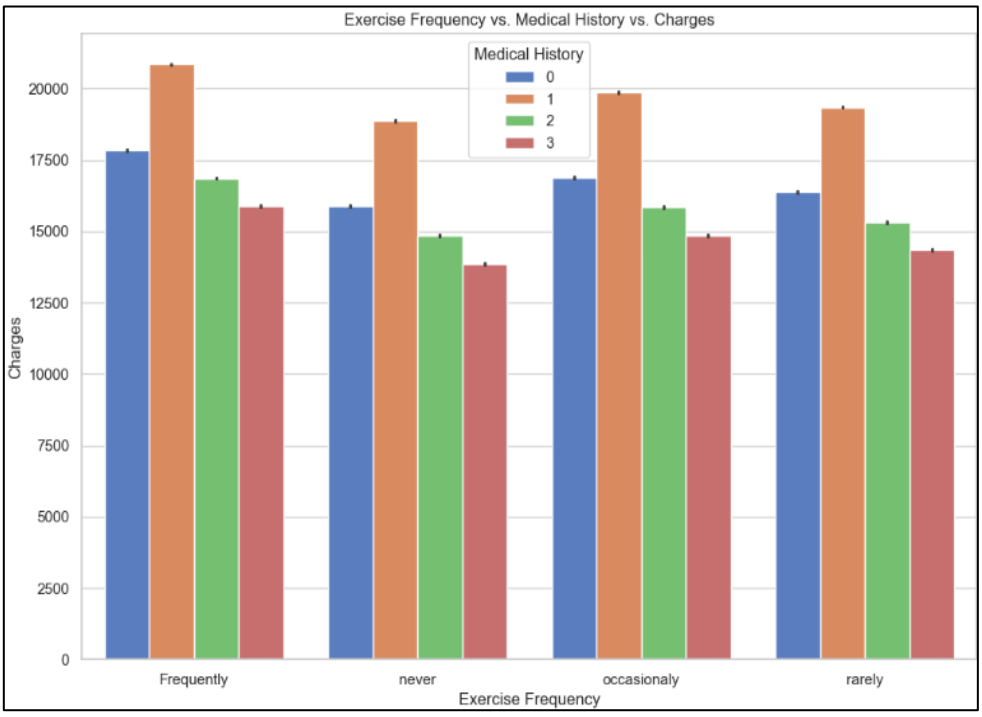
	df	sum_sq	mean_sq	F	PR(>F)
C(occupation)	3.0	6.208e+11	2.069e+11	10960.741	0.0
Residual	999996.0	1.888e+13	1.888e+07	NaN	NaN

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-996.616	0.001	-1028.1844	-965.0476	True
0	2	-1492.1644	0.001	-1523.7236	-1460.6052	True
0	3	501.2185	0.001	469.6199	532.8171	True
1	2	-495.5484	0.001	-527.0933	-464.0035	True
1	3	1497.8345	0.001	1466.2502	1529.4188	True
2	3	1993.3829	0.001	1961.8078	2024.958	True

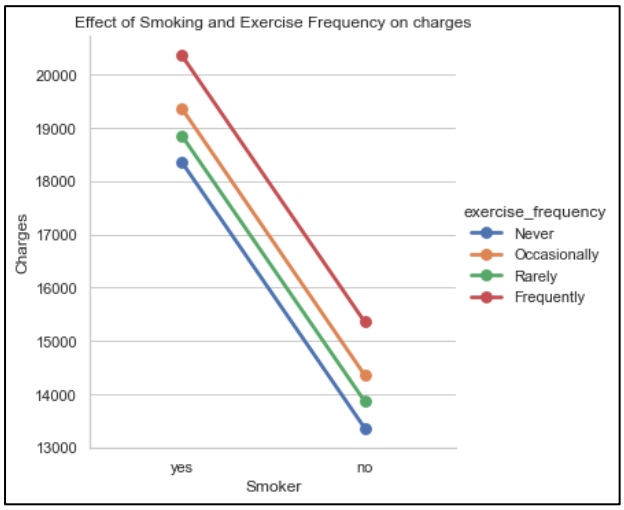
Charges according to Medical history and Exercise frequency together:

For the above plot categories are defined as follows: Diabetes (0), Heart Disease (1), High Blood Pressure (2), None (3). The bar plot above reveals a trend where individuals with a consistent exercise routine and a history of heart disease tend to experience higher charges. Subsequently, those with intermittent exercise habits and a history of heart disease also incur elevated charges. In contrast, individuals with neither exercise nor reported medical history have the lowest charges.



Charges according to Smoking and Exercise frequency together:

From the below plot, it is evident that People who frequently exercise and smoke tend to experience higher expenses. Conversely, individuals with no exercise routine and no smoking habit tend to have the lowest charges.



OLS Regression Results

Dep. Variable:	charges	R-squared:	0.349
Model:	OLS	Adj. R-squared:	0.349
Method:	Least Squares	F-statistic:	7.644e+04
Date:	Sun, 10 Dec 2023	Prob (F-statistic):	0.00
Time:	19:56:05	Log-Likelihood:	-9.5976e+06
No. Observations:	1000000	AIC:	1.920e+07
Df Residuals:	999992	BIC:	1.920e+07
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|
Intercept	1.535e+04	10.086	1522.357	0.000
C(exercise_frequency)[T.1]	-1993.4148	14.278	-139.610	0.000
C(exercise_frequency)[T.2]	-999.7571	14.267	-70.074	0.000
C(exercise_frequency)[T.3]	-1490.2606	14.235	-104.689	0.000
C(smoker)[T.1]:C(exercise_frequency)[0]	5005.5705	14.264	350.933	0.000
C(smoker)[T.1]:C(exercise_frequency)[1]	4995.6796	14.275	349.963	0.000
C(smoker)[T.1]:C(exercise_frequency)[2]	5010.7467	14.246	351.727	0.000
C(smoker)[T.1]:C(exercise_frequency)[3]	4989.9247	14.241	350.386	0.000

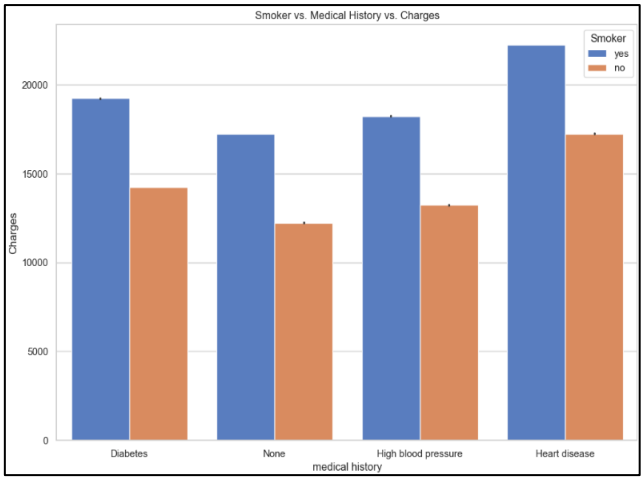
Omnibus:	16031.182	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14292.223
Skew:	0.243	Prob(JB):	0.00
Kurtosis:	2.672	Cond. No.	7.79

In the OLS regression findings, exercise frequency is denoted as occasionally (0), never (1), frequently (2), and rarely (3). The analysis highlights a pronounced impact on charges for individuals who both smoke and engage in frequent exercise, as indicated by the associated coefficients. This emphasizes the significant influence of this specific combination on the predicted charges.

```
Chi-square test for independence between smoker and exercise_frequency:
Chi2 value: 10.48006050411848
P-value: 0.01489677006840055
```

The chi-square test results indicate a statistically significant association between an individual's smoking status and their exercise frequency.

Charges according to Smoking and Medical History together:



	coef
Intercept	1.423e+04
C(medical_history)[T.1]	3002.9584
C(medical_history)[T.2]	-1002.4099
C(medical_history)[T.3]	-1995.5389
C(smoker)[T.1]:C(medical_history)[0]	5003.1979
C(smoker)[T.1]:C(medical_history)[1]	5004.8957
C(smoker)[T.1]:C(medical_history)[2]	4997.2858
C(smoker)[T.1]:C(medical_history)[3]	4993.3307

From the above it is evident that people with a history of heart disease who smoke tend to experience higher medical charges. In contrast, those without any medical history and who do not smoke tend to have the lowest incurred charges.

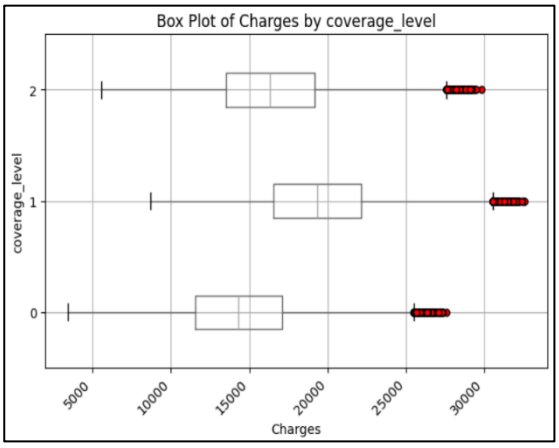
Charges by Coverage level:

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.217
Model:	OLS	Adj. R-squared:	0.217
Method:	Least Squares	F-statistic:	1.386e+05
Date:	Sun, 10 Dec 2023	Prob (F-statistic):	0.00
Time:	19:54:04	Log-Likelihood:	-9.6896e+06
No. Observations:	1000000	AIC:	1.938e+07
Df Residuals:	999997	BIC:	1.938e+07
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025
Intercept	1.439e+04	6.766	2127.430	0.000	1.44e+04
C(coverage_level)[T.1]	5008.7422	9.572	523.256	0.000	4989.981
C(coverage_level)[T.2]	2019.1344	9.568	211.020	0.000	2000.381

Omnibus:	18166.650	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12353.148
Skew:	0.155	Prob(JB):	0.00
Kurtosis:	2.553	Cond. No.	3.73

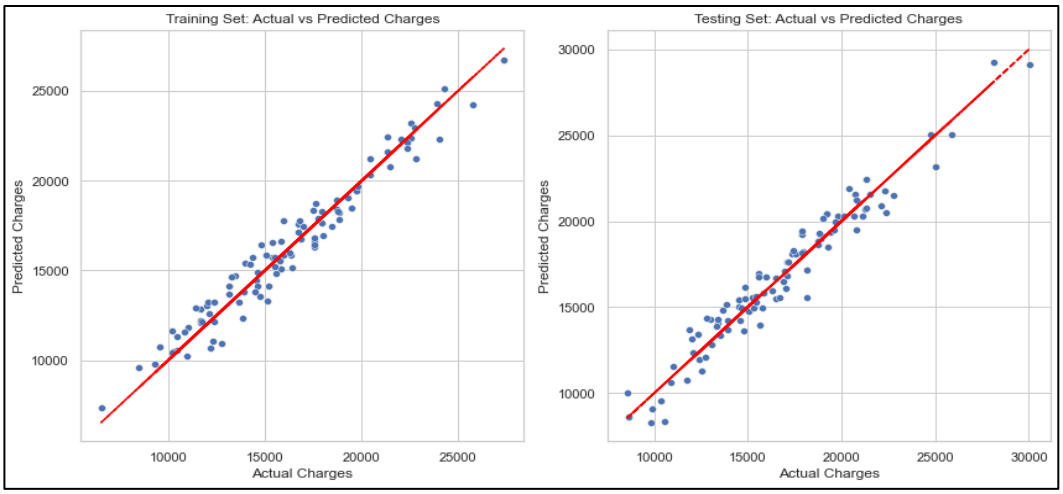


The classification includes designations such as basic (0), premium (1), and standard (2). The OLS regression findings, along with the information gleaned from the box plot, distinctly indicate that customers categorized as "Premium" bear the highest charges. Following this, "Standard" customers incur charges at an intermediate level, and "Basic" customers experience the lowest charges in a descending order of magnitude.

VI. Model Building

Linear Regression:

Linear Regression is one of the simplest and most used statistical techniques for predictive modeling. It assumes a linear relationship between the dependent variable and one or more independent variables. By fitting a linear equation to observed data, Linear Regression estimates the coefficients of the independent variables in the equation. The main goal is to find the best fit line that minimizes the differences between the predicted values and actual observations, which is often done through methods like least squares.



The graph above shows the predicted and actual values on a scatterplot which reflects the closeness of predicted values with the actual values of insurance premium using Linear Regression model.

Results without scaling and one hot encoding:

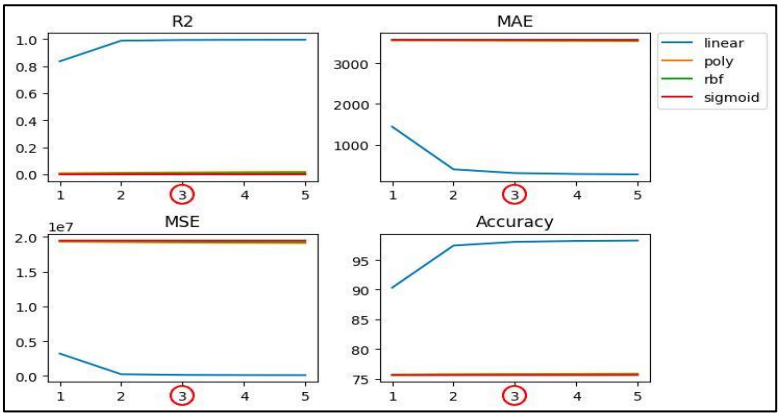
- Mean squared error: 9157484.05
- MAPE: 15.89 %
- Accuracy: 84.11 %.
- Train R2 score: 0.5290
- Test R2 score: 0.5294

Results with scaling and one hot encoding:

- Mean squared error: 839240.23
- MAPE: 4.85 %
- Accuracy: 95.15 %.
- Train R2 score: 0.9568
- Test R2 score: 0.9568

Support Vector Regression (SVR):

SVR is an extension of Support Vector Machines (SVM), a popular machine learning tool for classification problems. It works by mapping the input data into a high-dimensional feature space and then finding a hyperplane that best fits the data in this new feature space. The main idea is to minimize error, allowing for some errors to exist for certain points if they fall within a specified threshold.



The graph above shows four graphs, each comparing different metrics like R-squared, MAE, MSE, Accuracy for SVR model with different kernel and different regularization parameter C value. We can clearly see in the graph above the kernel Linear shows the best results and the C value after 3 has least change in graphs.

kernel: linear
C (regularization parameter): 3
gamma: scale

Results without scaling and one hot encoding:

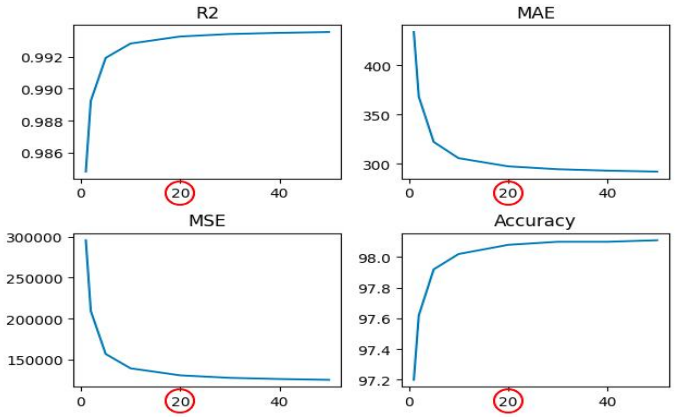
- Mean squared error: 9194809.95
- MAPE: 15.77 %
- Accuracy: 84.22 %.
- Train R2 score: 0.5272
- Test R2 score: 0.5275

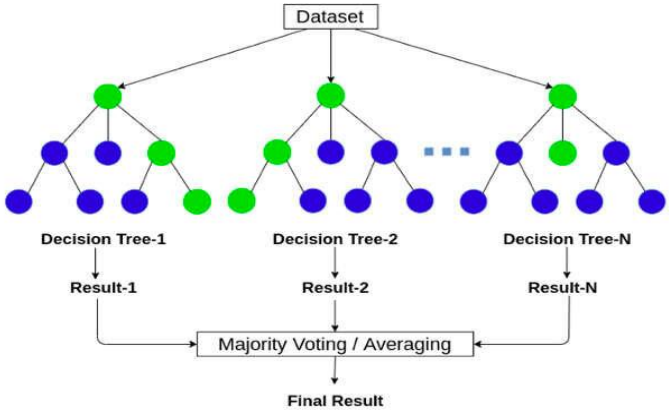
Results with scaling and one hot encoding:

- Mean squared error: 83623.51
- MAPE: 1.63 %
- Accuracy: 98.37 %.
- Train R2 score: 0.9955
- Test R2 score: 0.9948

Random Forest Regressor:

The Random Forest algorithm is a type of ensemble learning method, primarily used for classification and regression. A Random Forest Regressor builds multiple decision trees and merges them together to get a more accurate and stable prediction. Each tree in the forest is built from a sample drawn with replacement (bootstrap sample) from the training set. For regression tasks, the output of the Random Forest is the mean or average prediction of the individual trees.





The graph above shows the different metrics plotted on Y axis and the number of estimators on X axis of a line chart. We can see after 20 estimators there is very less change in the graph.

Results without scaling and one hot encoding:

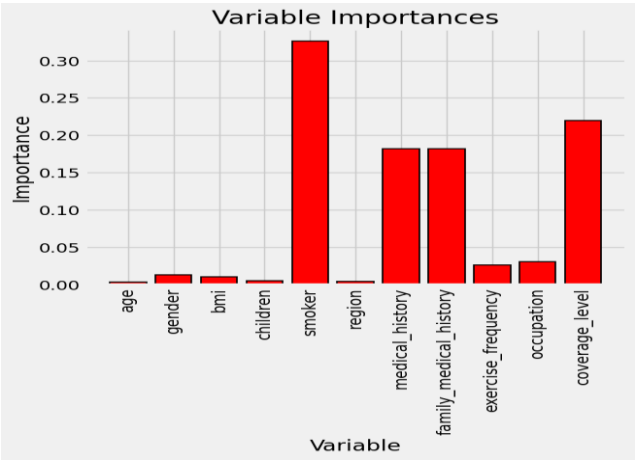
- Mean squared error: 137789.78
- MAPE : 1.96 %
- Accuracy: 98.03 %.
- Train R2 score: 0.9988
- Test R2 score: 0.9929

Results with scaling and one hot encoding:

- Mean squared error: 130913.70
- MAPE : 1.92 %
- Accuracy: 98.08 %.
- Train R2 score: 0.9989
- Test R2 score: 0.9932

Gradient Boosting:

Gradient Boosting is another ensemble technique that builds a model in a stage-wise fashion like other boosting methods but uses a different approach for training. It constructs new trees that predict the residuals or errors of prior trees and then combines these trees in a forward stage-wise manner. This method is known for its effectiveness in handling various types of data and its ability to improve prediction accuracy, but can be prone to overfitting if not properly tuned.



Params after Hypertuning:
GradientBoostingRegressor (n_estimators=25, learning_rate=0.8, subsample= 0.9, max_depth=9, random_state=1)

Results without scaling and one hot encoding:

- Mean squared error: 126511.28
- MAPE : 1.89 %
- Accuracy: 98.10 %.
- Train R2 score: 0.9941
- Test R2 score: 0.9935

Results with scaling and one hot encoding:

- Mean squared error: 133939.26
- MAPE : 1.92 %
- Accuracy: 98.07 %.
- Train R2 score: 0.9969
- Test R2 score: 0.9931

XGBoost:

XGBoost stands for Extreme Gradient Boosting and is an efficient implementation of the Gradient Boosting framework. This algorithm has gained a lot of popularity in machine learning competitions for its performance and speed. XGBoost provides a parallel tree boosting technique, which efficiently solves many data science problems in a fast and accurate way. It includes features like handling missing values, regularizing to avoid overfitting, and cross-validation at each iteration of the boosting process.

Params after Hypertuning:
XGBRegressor (n_estimators=200, gamma=0.3, max_depth=4, random_state=1)

Results without scaling and one hot encoding:

- Mean squared error: 85780.12
- MAPE: 1.63
- Accuracy: 98.37 %.
- Train R2 score: 0.9957
- Test R2 score: 0.9956

Results with scaling and one hot encoding:

- Mean squared error: 85635.95
- MAPE: 1.63
- Accuracy: 98.37 %.
- Train R2 score: 0.9958
- Test R2 score: 0.9956

Results of All Models:

Regression Model	M.S.E	M.A.P.E	ACCURACY	Train_R2 Score	Test_R2 Score
Linear	**839240.24**	**4.85**	**95.15%**	**0.9568**	**0.9568**
Random Forest	**130913.70**	**1.92**	**98.08%**	**0.9989**	**0.9932**
SVR	**83623.51**	**1.63**	**98.37%**	**0.9955**	**0.9948**
Gradient Boosting	**133939.26**	**1.92**	**98.07%**	**0.9969**	**0.9931**
XG Boost	**85635.95**	**1.63**	**98.37%**	**0.9958**	**0.9956**

1. Linear Regression: It has a relatively high Mean Squared Error (M.S.E) of 839240.24, indicating that the predictions may have significant variance from the actual values. The Mean Absolute Percentage Error (M.A.P.E) is 4.85, suggesting that on average, the prediction is off by 4.85%. The accuracy is quite high at 95.15%. Both the training and test R-squared scores are 0.9568, which means the model explains 95.68% of the variability in the response variable for both the training and test sets.

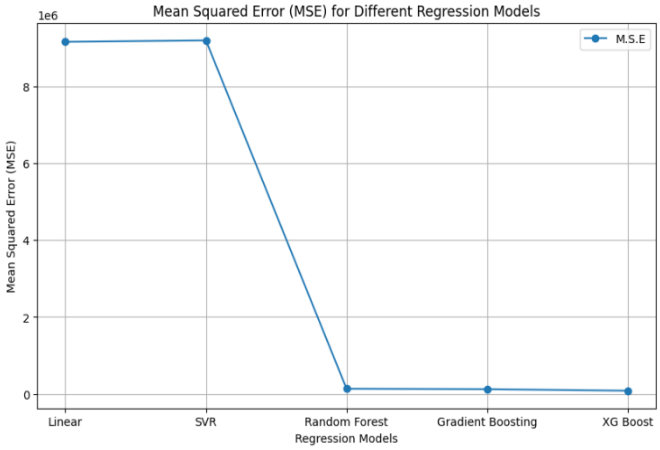
2. Random Forest: This model shows a much lower M.S.E of 130913.70 and a lower M.A.P.E of 1.92, which are improvements over the Linear Regression model. The accuracy is higher at 98.08%. The R-squared scores are very high (0.9989 for training and 0.9932 for test), indicating the model very accurately fits the data.

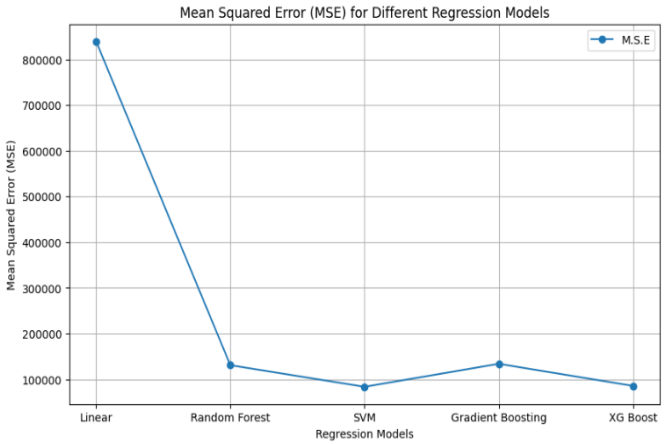
3. SVR (Support Vector Regression): The SVR model has an M.S.E of 83623.51and an M.A.P.E of 1.63, which are the lowest errors among all models listed. The accuracy is the highest at 98.37%. The R-squared scores are 0.9955 for training and 0.9948 for test, suggesting excellent model performance.

4. Gradient Boosting: This model has an M.S.E of 133939.26 and an M.A.P.E of 1.92. The accuracy is slightly lower than SVR at 98.07%. The R-squared scores are 0.9969 for training and 0.9931 for test, indicating a strong fit to the data.

5. XG Boost: The XG Boost model posts an M.S.E of 85635.95 and the lowest M.A.P.E, tied with SVR, at 1.63. It also has the highest shared accuracy with SVR at 98.37%. The R-squared scores are slightly higher than SVR, with 0.9958 for training and 0.9956 for test, showing excellent fit and predictive capability.

Overall, SVR and XG Boost are top performers across these metrics, with Random Forest also showing strong results, particularly in terms of R-squared scores. Linear Regression, while still performing reasonably well, lags behind the others in terms of these specific metrics.





- Scaling and one-hot encoding play a crucial role in enhancing the performance of linear regression and Support Vector Regression (SVR).

- Random Forests exhibit resilience to the absence of scaling and one-hot encoding, showcasing their ability to handle different scales and categorical variables effectively.
- While scaling and encoding may offer some benefits to Gradient Boosting models (GB and XGB), the extent of improvement is likely to be less significant compared to linear models.
- Overall, tree models consistently outperformed in both scenarios with and without preprocessing steps.

Conclusion

Through this project, we conducted a thorough analysis of the significant factors influencing insurance charges using various machine-learning approaches. Our implementation of five machine learning models resulted in accurate predictions of charges. Our research identified smoking, medical history, family medical history, and coverage level as the primary influencing factors on insurance costs. The objective of this project is to optimize the entire process for both consumers and insurance providers. This optimization aims to directly benefit consumers by ensuring they receive the best prices and providing insurance providers with the maximum turnout from leads, potentially leading to more insurance policies. Furthermore, this solution can seamlessly integrate with various applications such as SAP or CRM, catering to real-time business requirements.

Reference

[1] Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health*, *19*(13), 7898.