HEALTH INSURANCE PREDICTION

MENTOR :

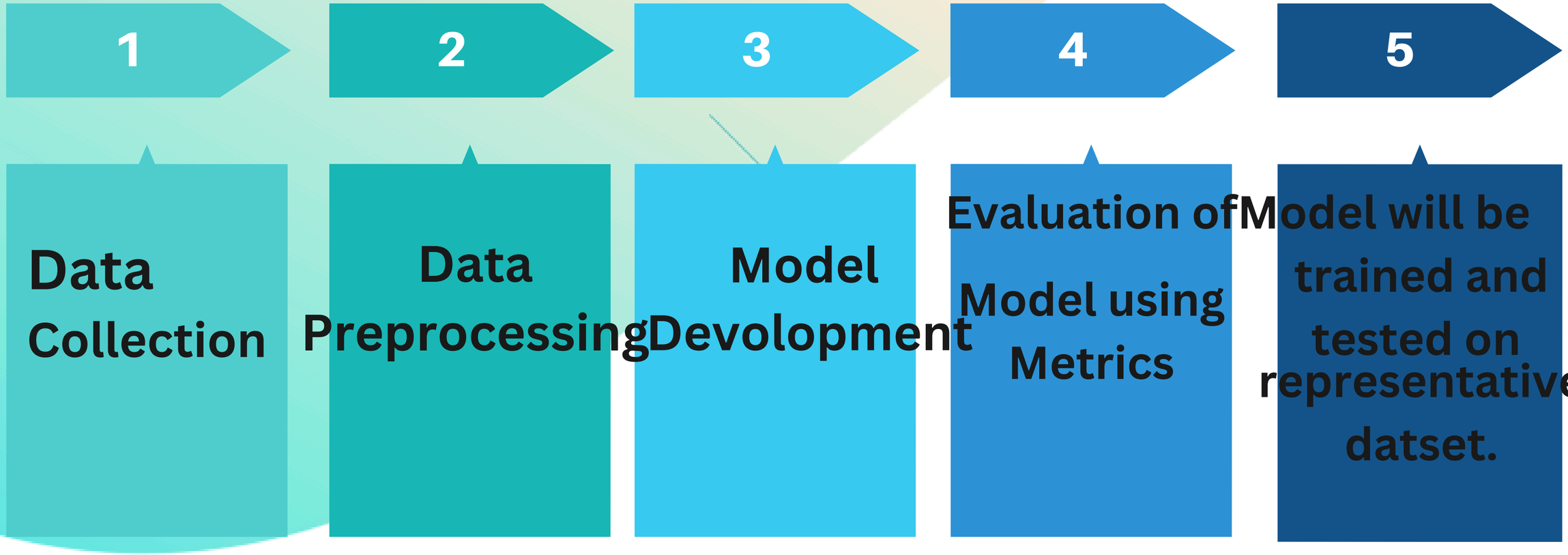
T.SUJANAVAN

K.SHASHANK 2451-22-748-022
J.RITHIKA 2451-22-748-046
B.SATHVIKA 2451-22-748-009

PROBLEM STATEMENT

- This project aims to develop a predictive model using Gradient Boosting Regression to estimate health insurance premium based on client information and their medical history .
- Predictive health insurance premium models helps clients understand their premiums better and plan finances wisely.
- While it helps insurance companies in setting fair prices and improve operational efficiency

MODEL SELECTION STEPS



USERS

INSURANCE COMPANIES :
LOOKING TO SET ACCURATE PREMIUM RATES .

POLICY HOLDERS :
SEEKING TO ESTIMATE THEIR COST .

TOOLS AND TECHNOLOGIES USED :

TOOLS

1. LANGUAGE :
   - Python

LIBRARIES AND FRAMEWORK :
- Scikit -learn
- Pandas
- Numpy

3. DATA COLLECTION
   - Pandas
   - Kaggle

4. DATA VISUALIZATION :
   - Matplot -lib
   - Seaborn

5. MACHINE LEARNING MODEL :
   - Linear Regression
   - Gradient Boosting Regressor

TOOLS

6. EVALUATION METRICS :

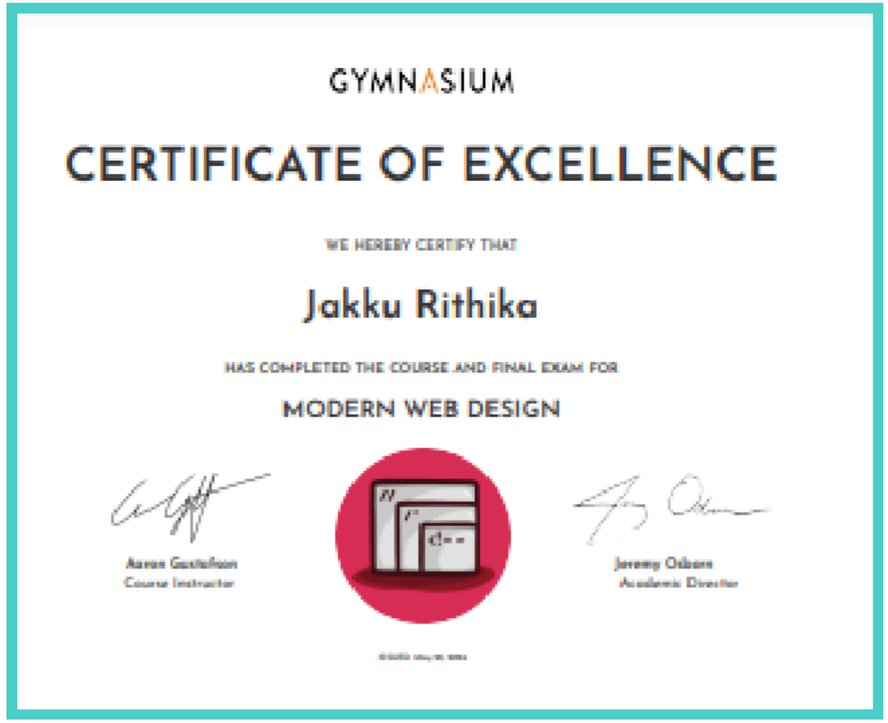
- MSQ(mean squared error)
- MAS(mean absolute error)
- r2_score

7.FRAME WORK

- Streamlit

CERTIFICATION :

From GYMNASIUM



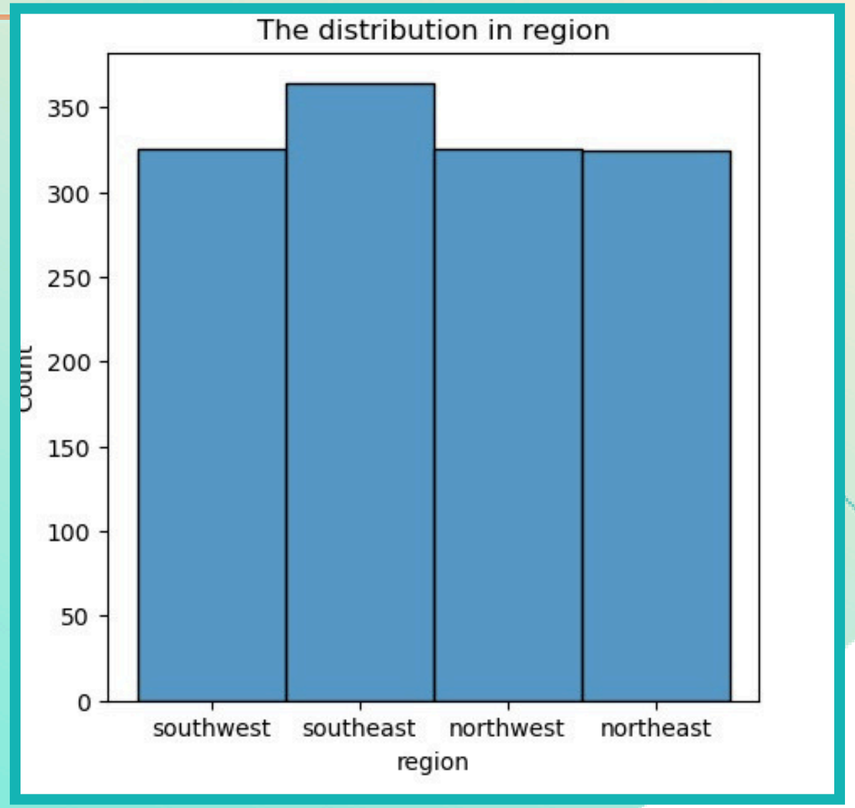
DATA COLLECTION

- The dataset was taken from Kaggle.
- Data set includes the following dependent variable Premium Price and all other are independent variables.

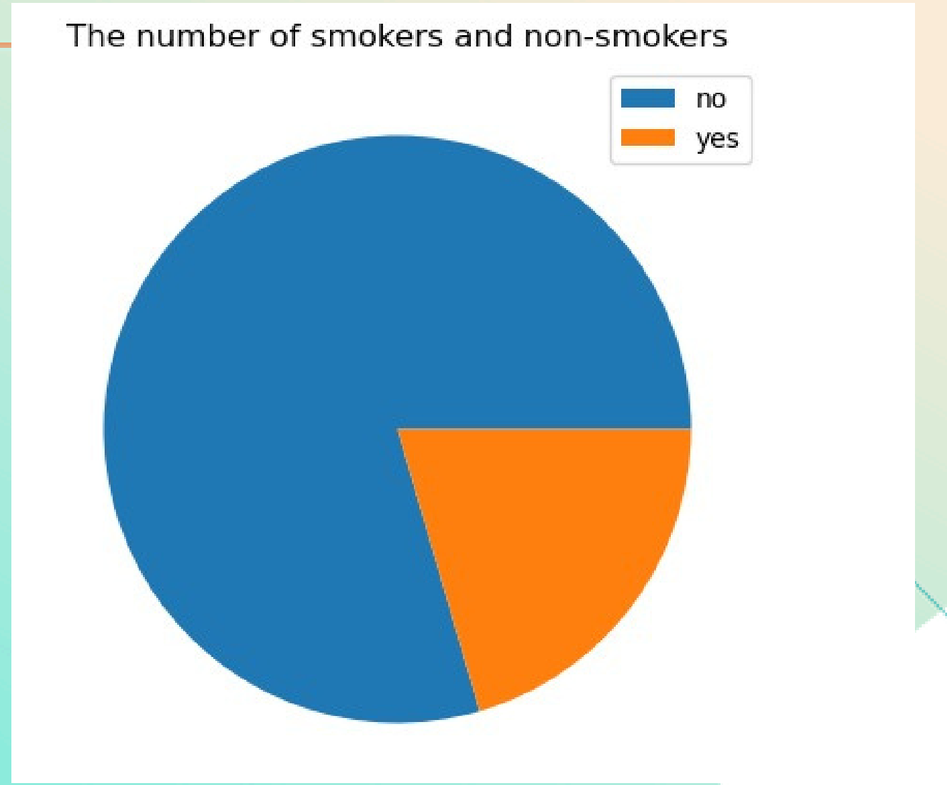
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

EDA(Exploratory Data Analysis)

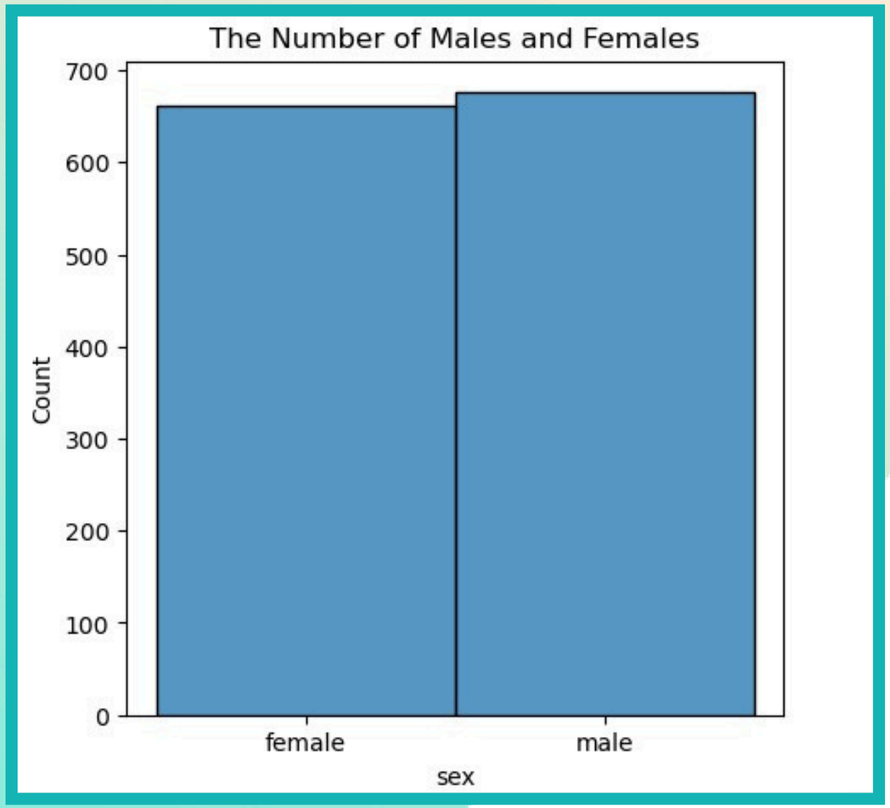
- Understanding how data is organized.
- Checking relationships between factors like age, weight, height etc..
- Removing Duplicates
- Impute Missing values
- Converting categorical data into numerical data



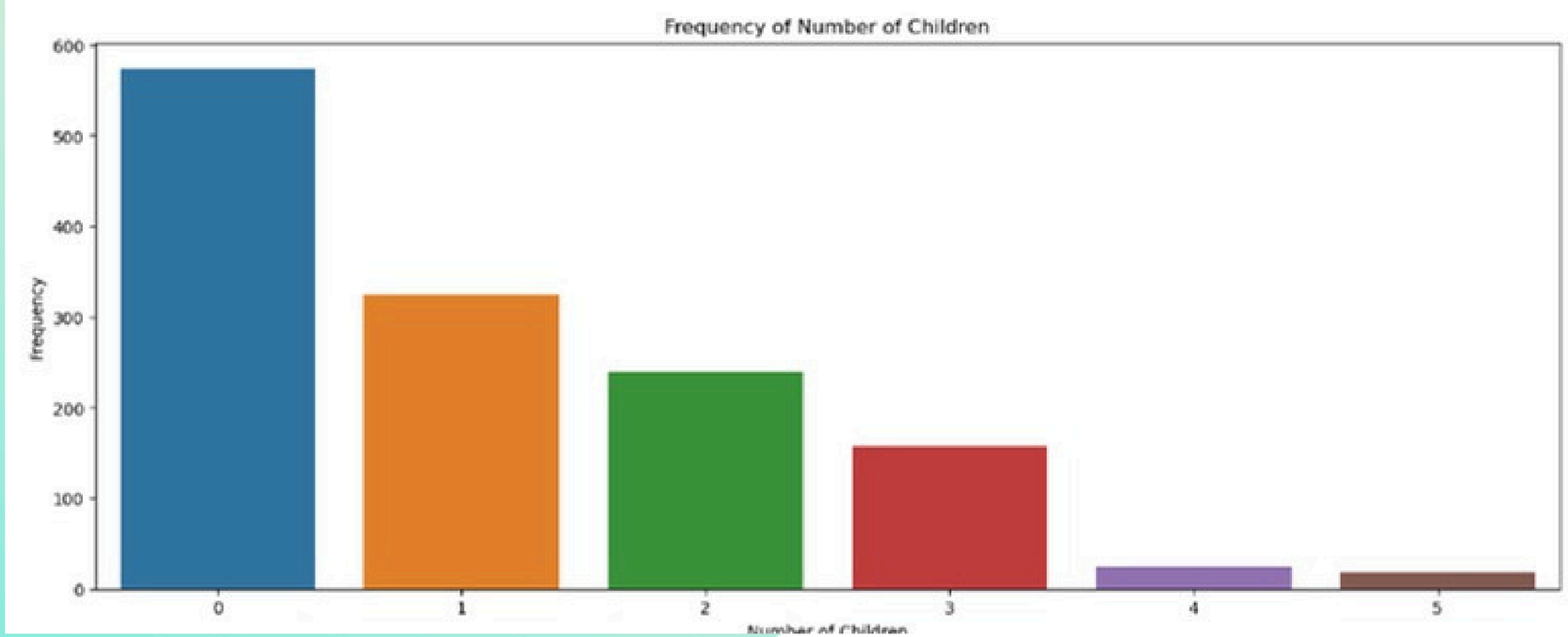
"The distribution of the 'region' variable reveals that the majority of data points are concentrated in the southeast region, with fewer data points in the northwest and southwest regions.



"The graph showing the number of smokers versus non-smokers reveals that the dataset is predominantly composed of non-smokers, with non-smokers outnumbering smokers by a ratio of approximately 3:1.



"The graph showing the number of females versus males reveals that the dataset is predominantly composed of females, with females outnumbering males by a ratio of approximately 2:1.



"The bar plot showing the frequency of the number of children reveals that the majority of individuals in the dataset have 0 to 2 children. There are fewer individuals with 3 or more children.

Data splitting into Training data and Testing data :

```
# Splitting Training and Testing Dataset

xtrain , xtest , ytrain , ytest = train_test_split(x , y , test_size = 0.20 , random_state = 42)
```

- Training_data size=80%
- Testing_data size=20%

- Splitting is done using train_test_split from sklearn library

```
# Building Model

mod = LinearRegression()
mod.fit(xtrain , ytrain)
```

```
LinearRegression
LinearRegression()
```

Linear Regression model

```
# Metrics

print("The training accuracy is " , mod.score(xtrain , ytrain)*100 , "%")
print("The testing accuracy is " , mod.score(xtest , ytest)*100 , "%")
print("The mean absolute error is " , mean_absolute_error(pred , ytest))
print("The mean squared error is ", mean_squared_error(pred , ytest))
```

```
The training accuracy is  72.9449036210828 %
The testing accuracy is  80.61028038524825 %
The mean absolute error is  4184.992650402236
The mean squared error is  35629785.59267284
```

accuracy=72%

Gradient Boosting Regressor

```
# Model Building : gradient boosting Regression
from sklearn.ensemble import GradientBoostingRegressor
# Initialize and train the model
GBR = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)
GBR.fit(xtrain, ytrain)

# Predict on test set
ypred_2 = GBR.predict(xtest)
```

```
from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score
mae = mean_absolute_error(ypred_2, ytest)
mse = mean_squared_error(ypred_2, ytest)
r2 = r2_score(ytest, ypred_2)
print('Mean absolute error :',mae)
print('Mean squared error :',r2)
print('R^2 :', r2)
```

```
Mean absolute error : 2434.3001347793856
Mean squared error : 0.9034737694301922
R^2 : 0.9034737694301922
```

accuracy=90%

EVALUATION OF MODEL

- Evaluation of model is done using metrics such as :

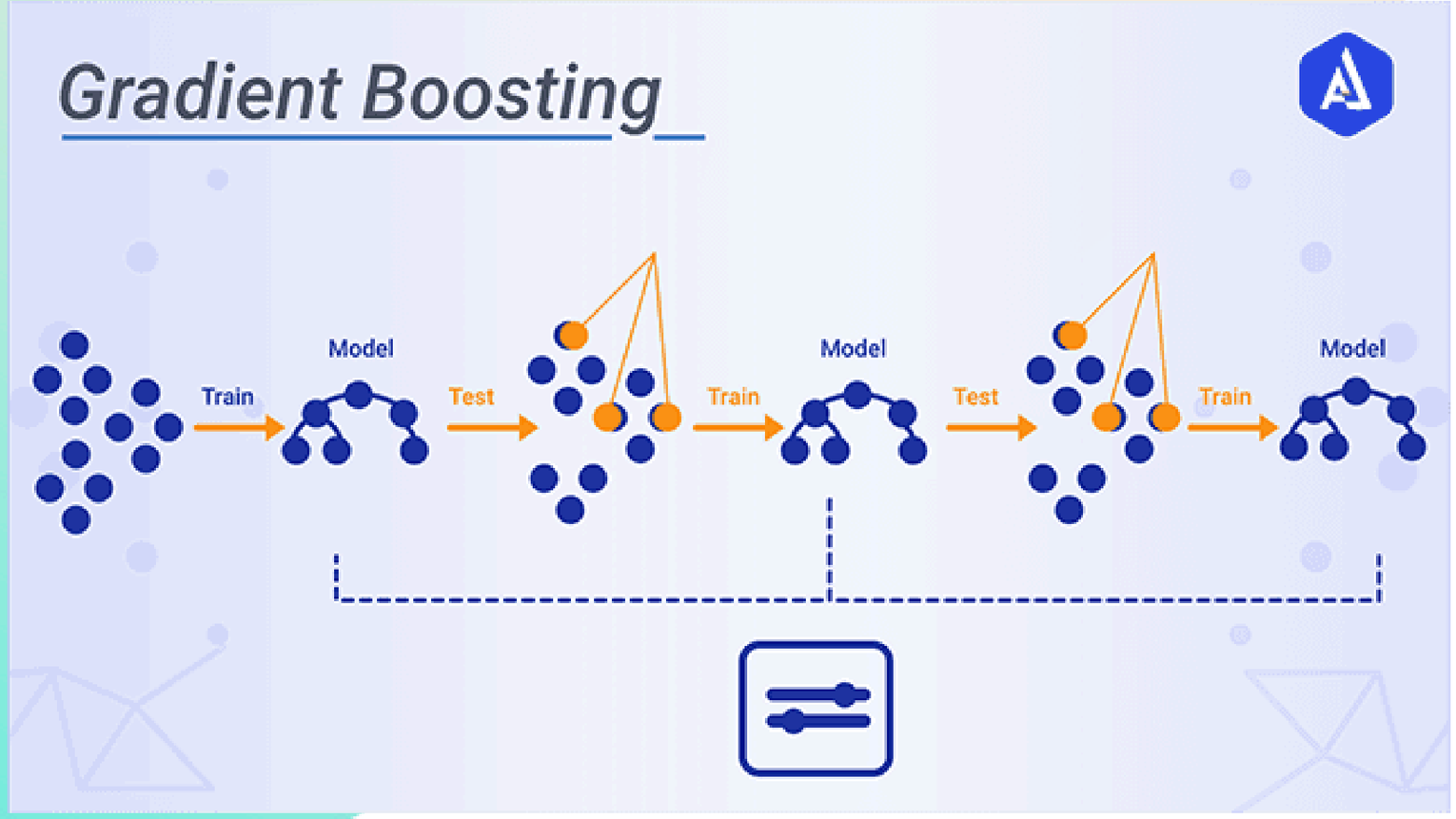
    1. MAE(Mean absolute error)
    2. MSE(Mean squared error)
    3. r2_score

- After evaluation of model performance we have found Gradient Boosting is giving more accurate values.

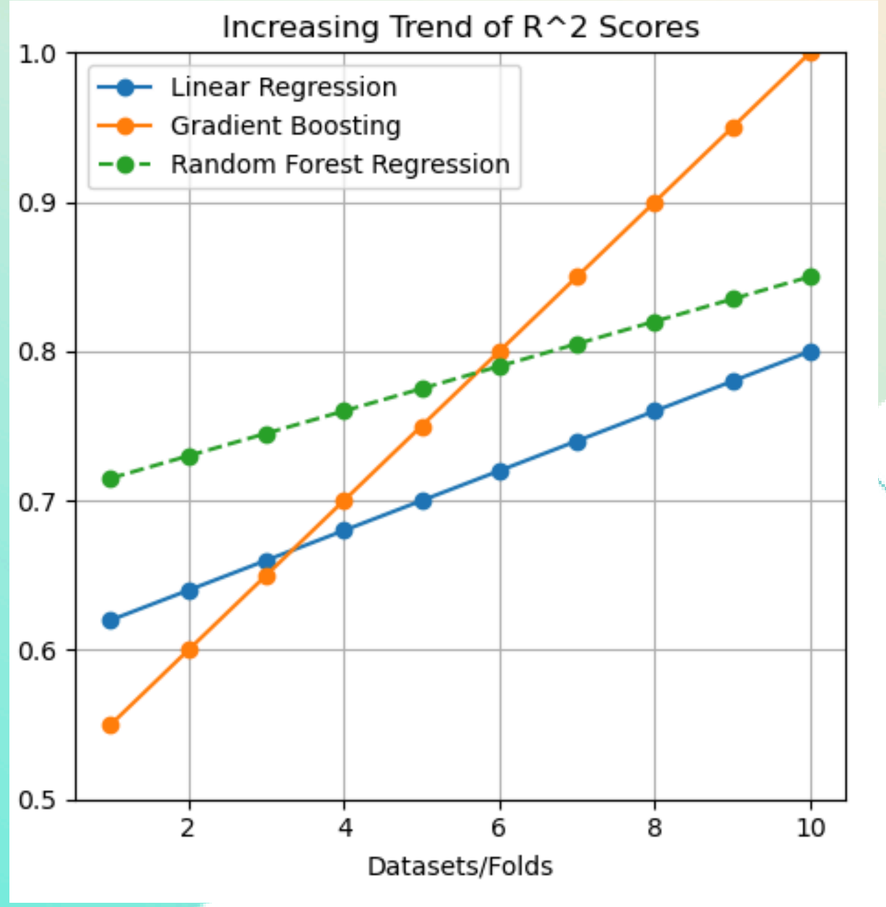
The algorithm used in this code is Gradient Boosting Regression. This algorithm is implemented using the GradientBoostingRegressor class from the sklearn.ensemble module in scikit-learn.

Gradient Boosting is an ensemble machine learning technique that builds a predictive model in a stage-wise fashion from an ensemble of weak learners, typically decision trees. It combines the predictions of several base estimators in order to improve robustness over a single estimator. This technique is well-suited for regression and classification problems.

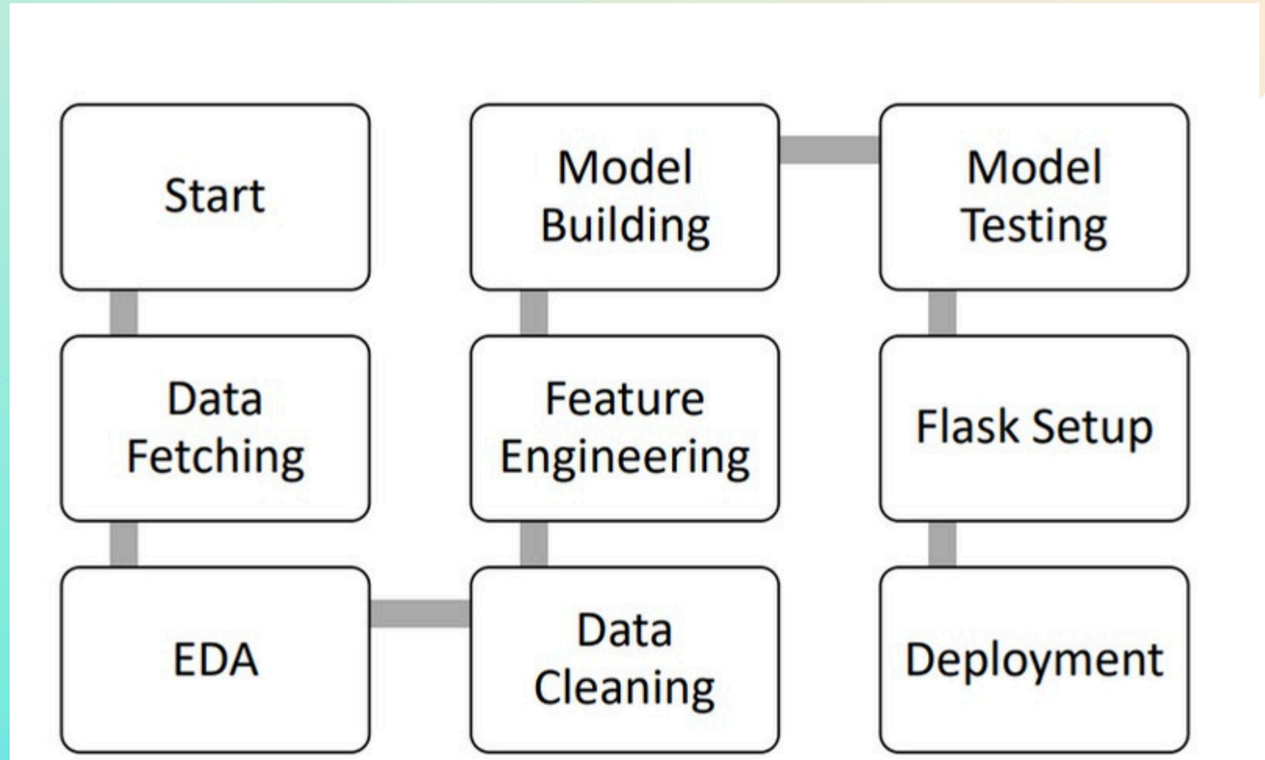
In the provided code, GradientBoostingRegressor is used to predict medical insurance costs based on various features such as age, gender, BMI, number of children, smoking status, and region.


Gradient Boosting
Train
Model
Test
Train
Model
Test
Train
Model

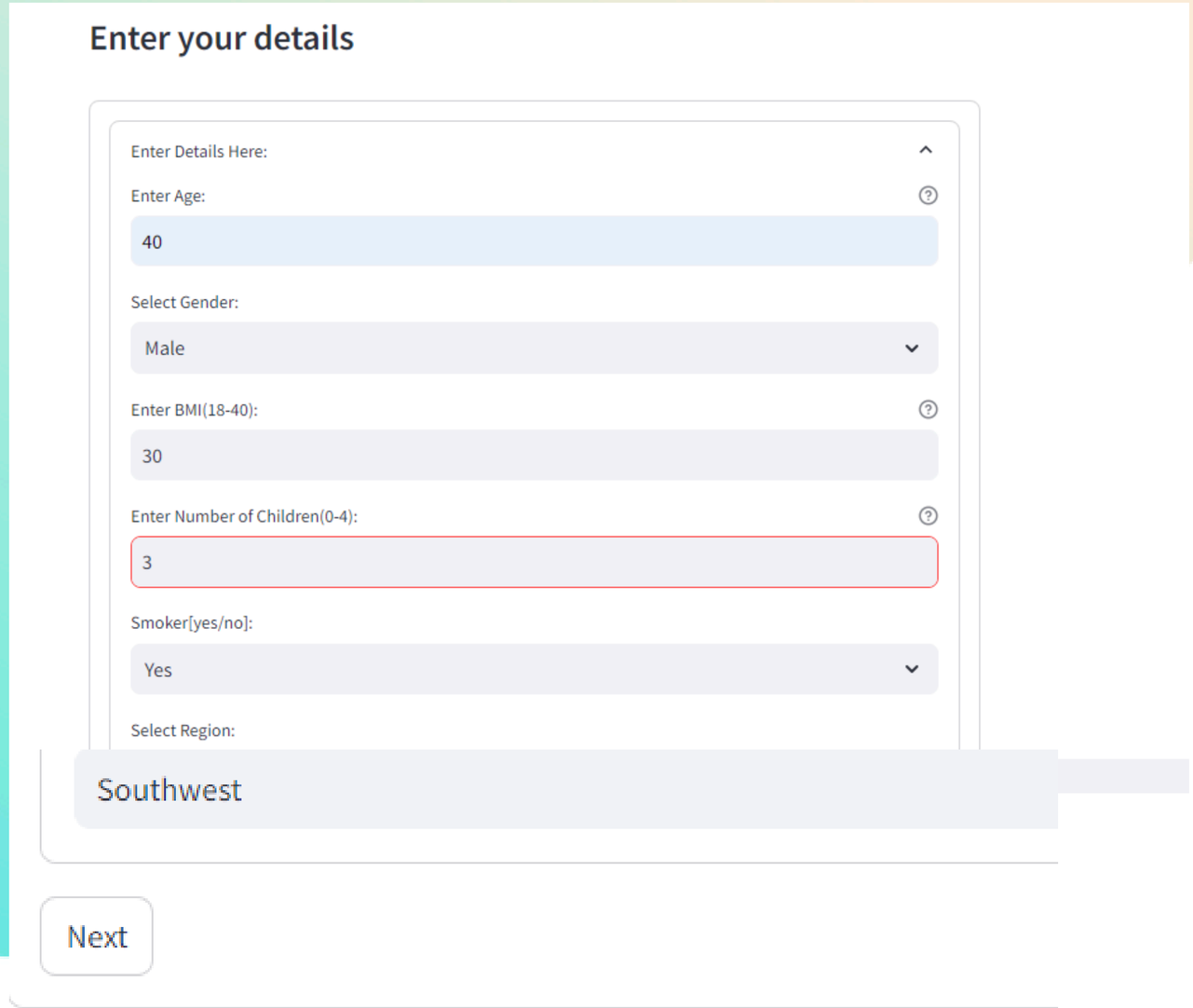
Accuracy comparison graph between 3 algorithms



ARCHITECTURE



RESULT :


Enter your details
Enter Details Here:
Enter Age:
40
Select Gender:
Male
Enter BMI(18-40):
30
Enter Number of Children(0-4):
3
Smoker[yes/no]:
Yes
Select Region:
Southwest
Next

RESULT :

Your Prediction

Your Predicted Health Insurance Price is: 36131

CONCLUSION

- Gradient Boosting regression proved to be an effective model for predicting health insurance premiums with high accuracy.

- These predictions can help insurance companies and policy holders in pricing strategies and risk assessment, improving overall efficiency.

REFERENCES

[1] Medical Insurance Raj dataset, https://www.kaggle.com/datasets/rajgupta2019/medical-insurance-dataset

[2] Introduction Gupta, to Streamlit, Great Learning 2019, Academy, https://www.mygreatlearning.com/academy/learn-for-free/courses/introduction-to streamlit

[3] Modern Web Design, The Gymnasium, https://thegymnasium.com/courses/course v1:GYM+107+0/about

[4] Introduction to Machine Learning, Ethem Alpaydın, ISBN: 9780262043793, March 17, 2020, The MIT Press

[5] Machine Learning, Tom M. Mitchell, 2013, McGraw Hill.

[6] Data Thinkers, 2021, 3. Project 2 Health Insurance Cost Prediction End To End Machine Learning Project https://www.youtube.com/watch?v=eshMzk8L3ic

Thank You