PROJECT OWNER:

AMADO DE JESÚS VÁZQUEZ ACUÑAS

# INSURENCE

BY: İCLAL ŞEYMA KOCA

# INSURENCE

THE PROJECT AIMS TO DETERMINE THE INSURANCE PRICE FOR THIS PERSON, CONSIDERING DIFFERENT PARAMETERS SUCH AS PEOPLE'S AGE, BODY MASS INDEX, ETC.

# LIBRARIES USED FOR THE PROJECT

Numpy

Shap

Seaborn

Pandas

Matplotlib

Numpy

Warnings

# CONTENT

EDA

FEATURE ENGINEERING

SELECTION OF MODEL IDEAL

DEFINITIVE MODEL

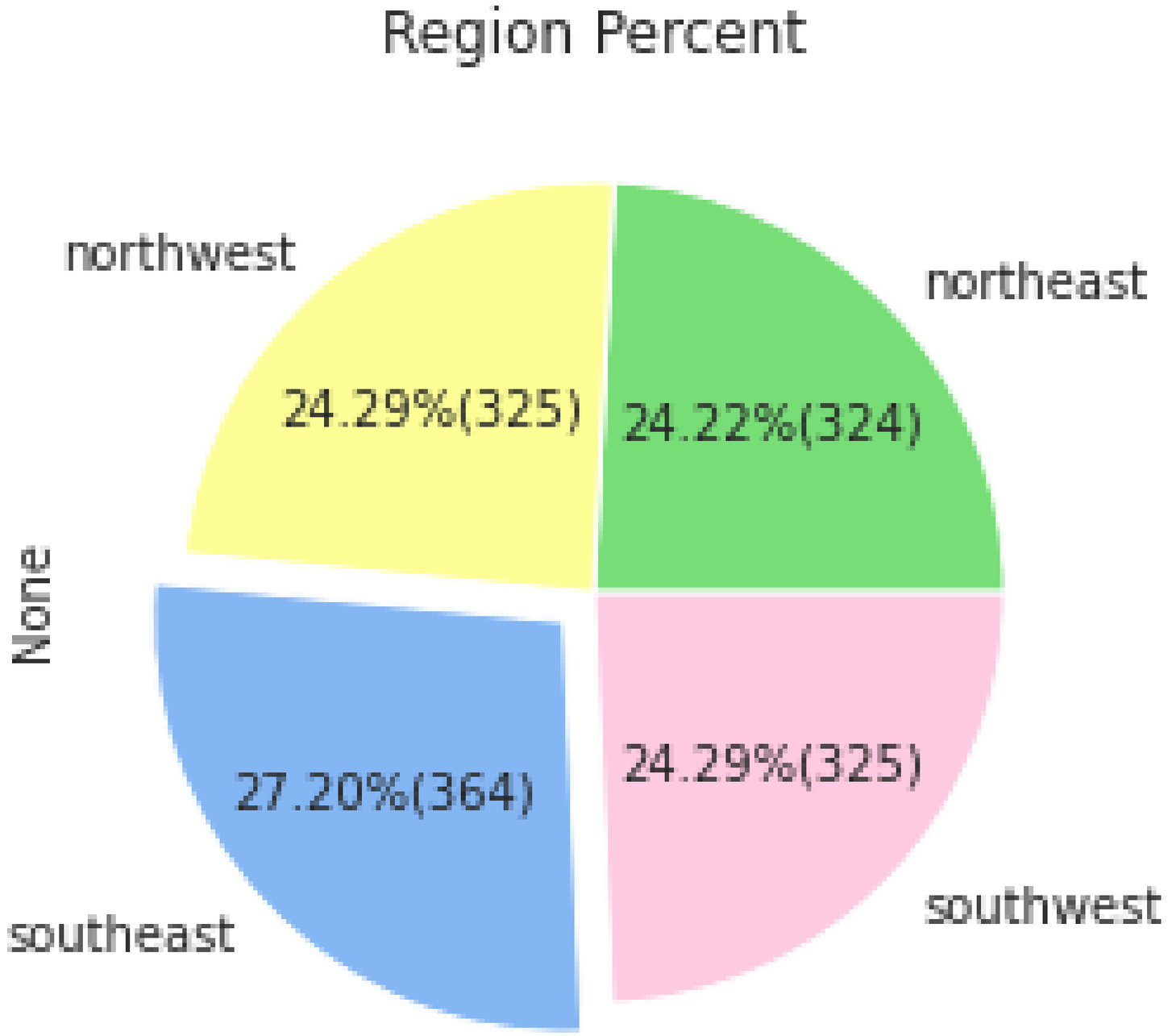
# 1.EDA

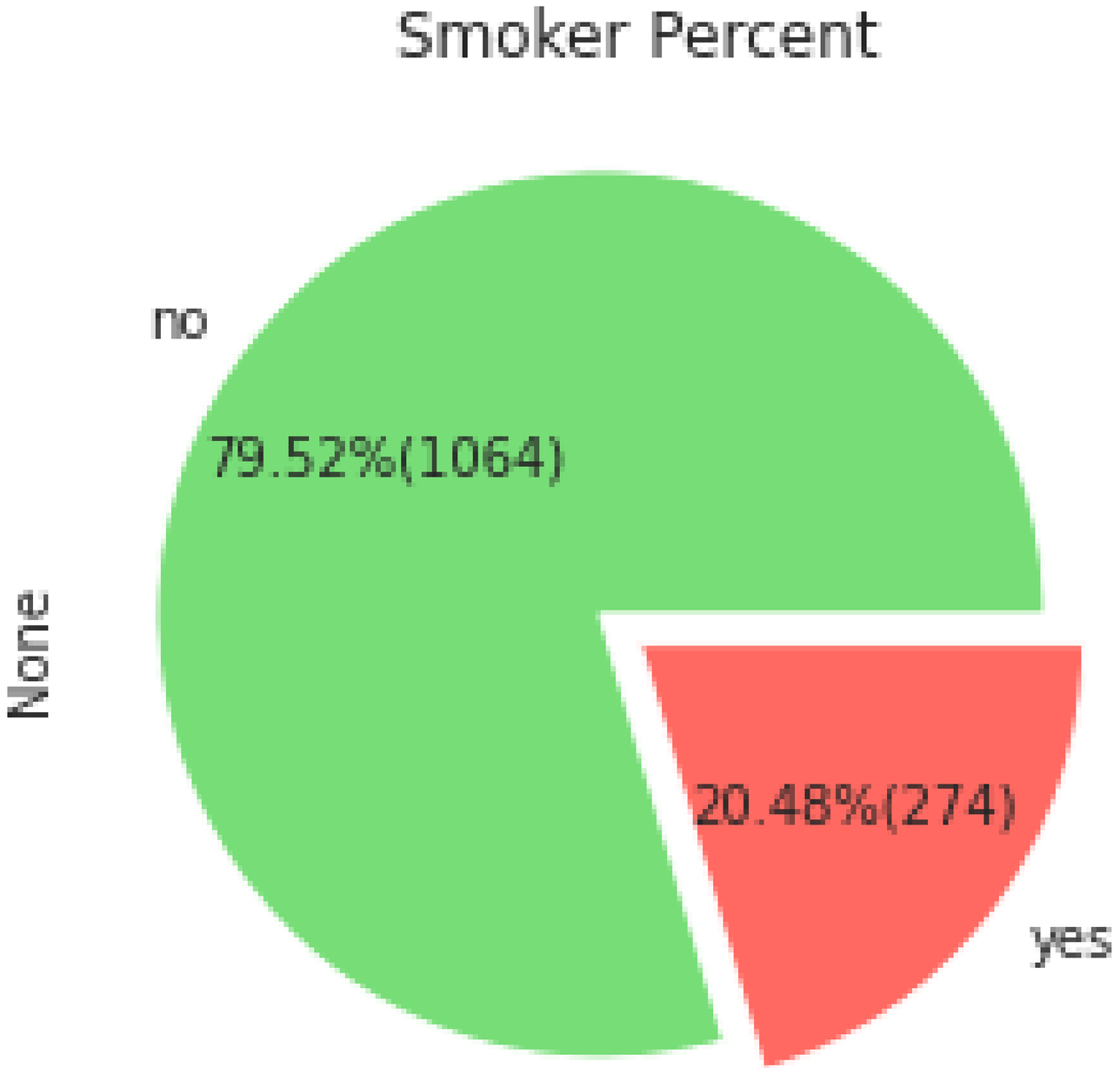
## IMPORTANT HEADINGS IN DATA

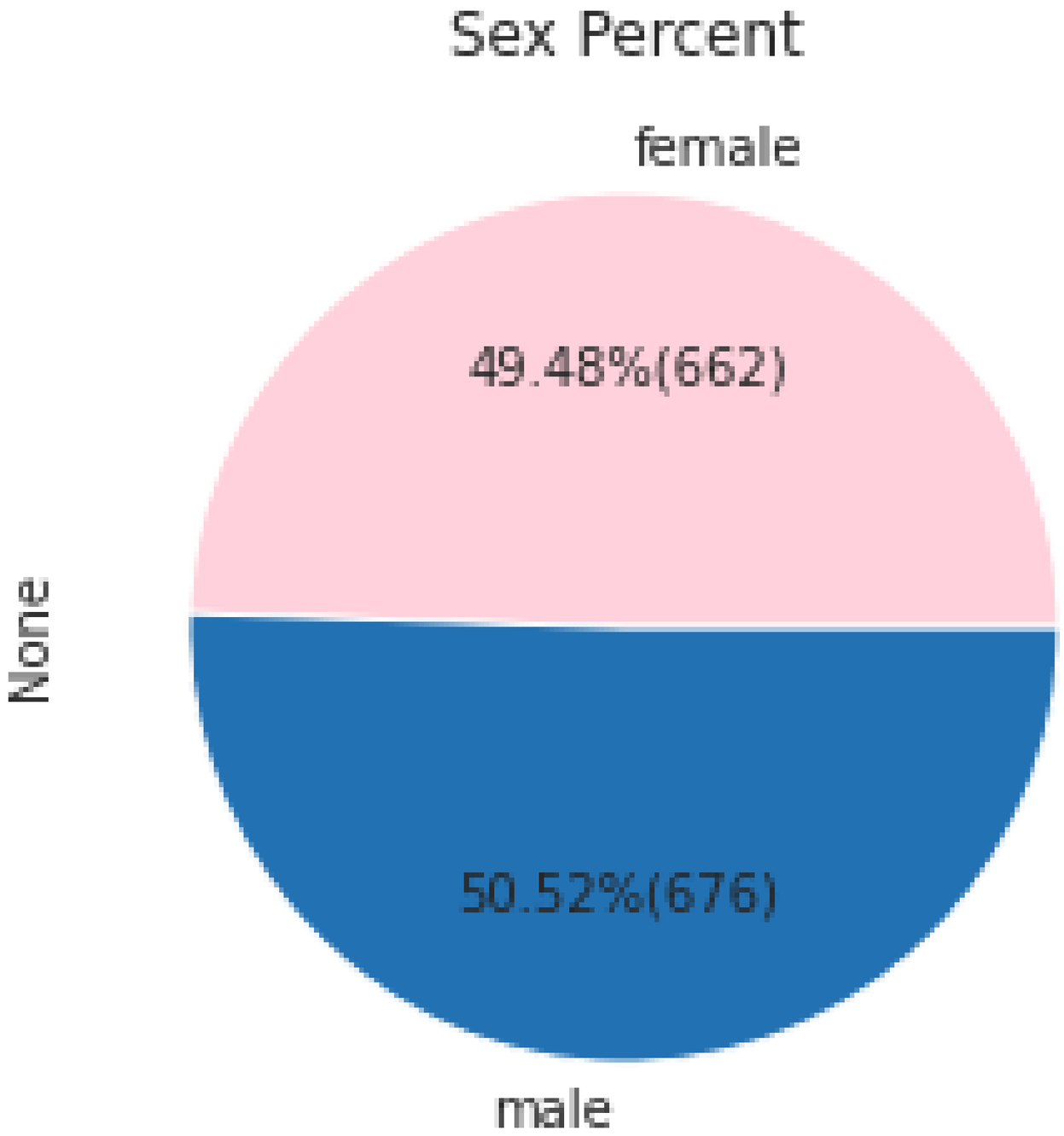
Age
Sex
BMI Body Mass Index
Children
Smoker
Region
Charges
Medical Problem

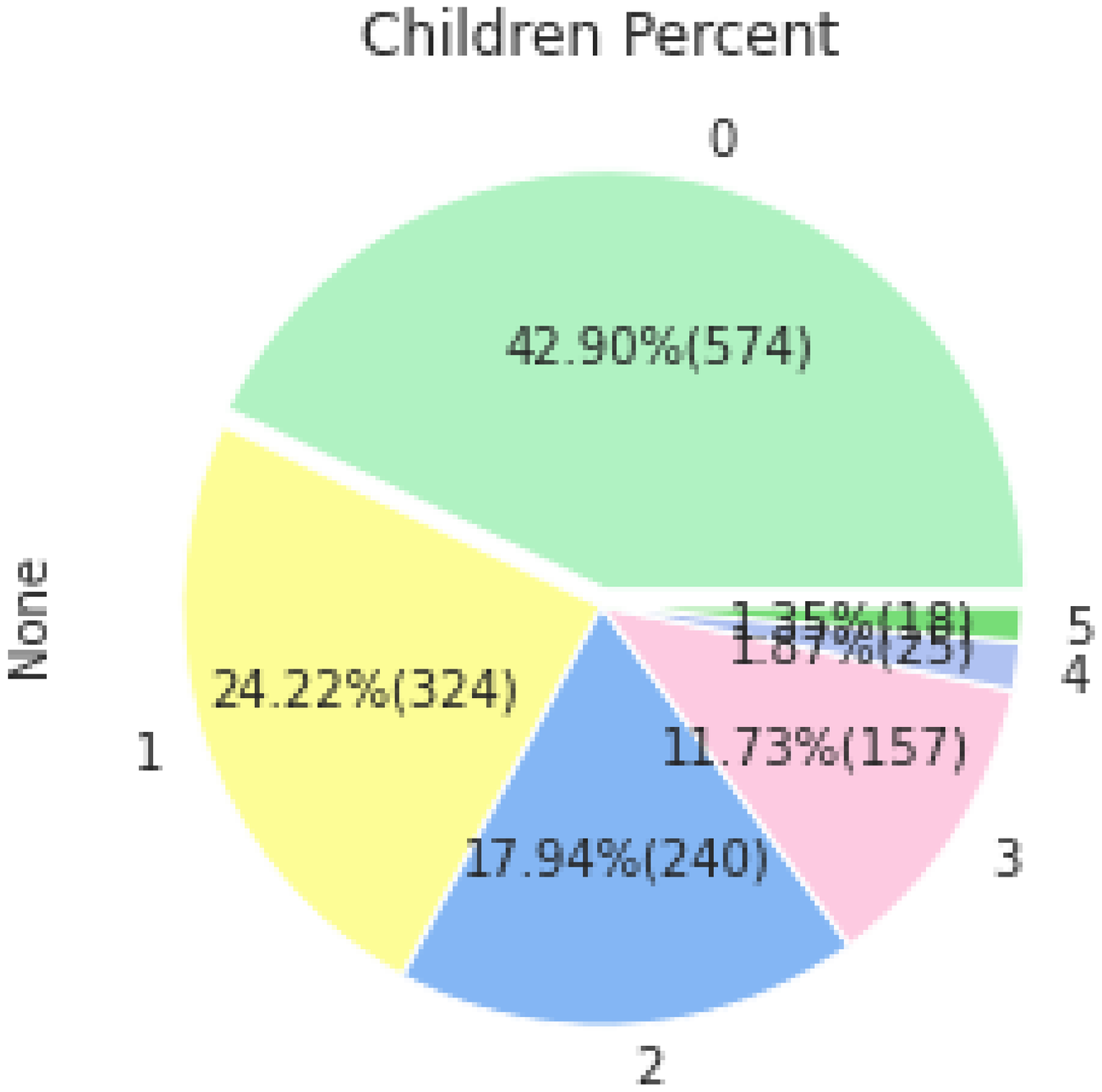
# EXPLORATORY DATA ANALYSİS (EDA)

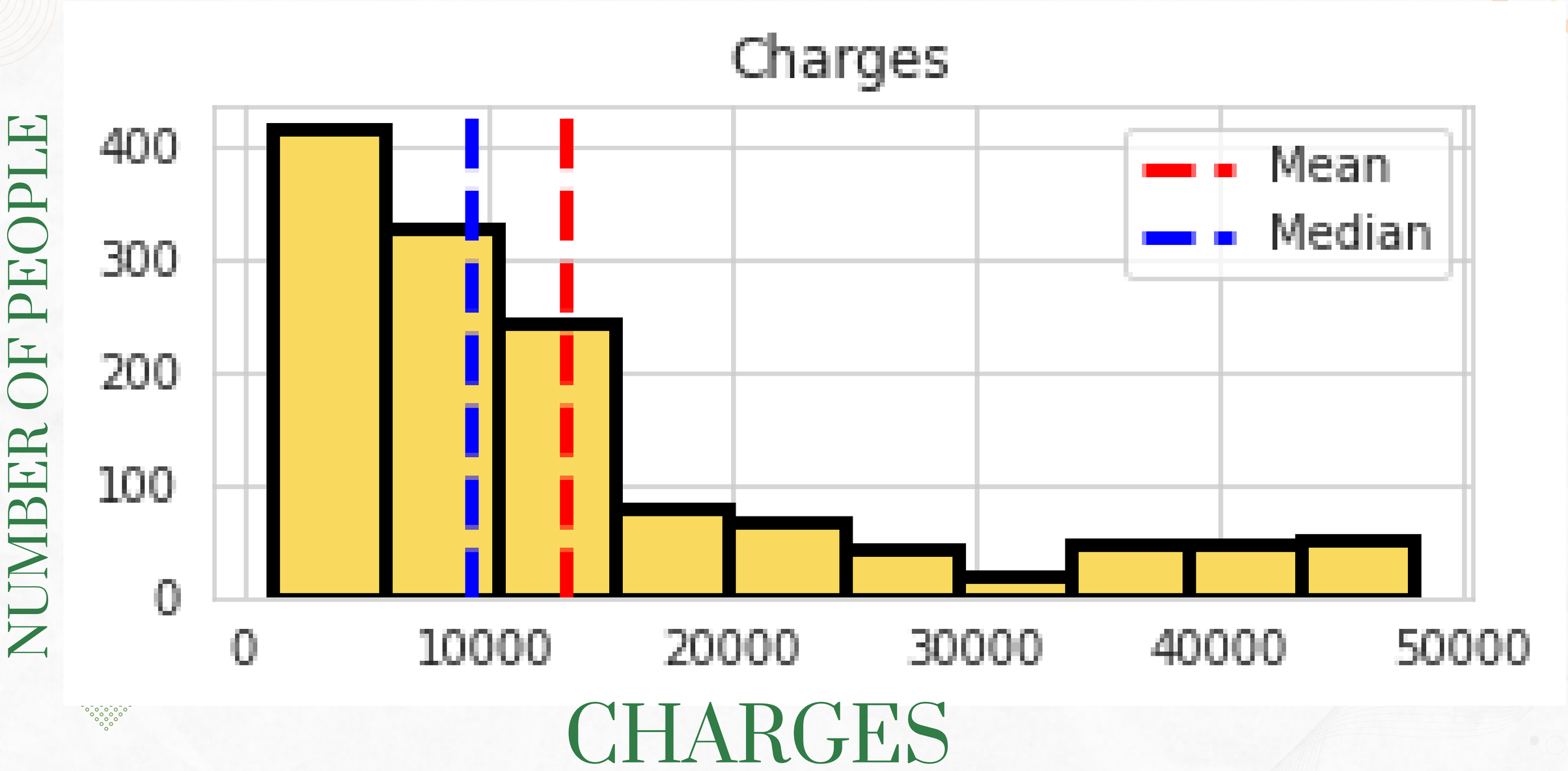
Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of summary statistics and graphical representations.

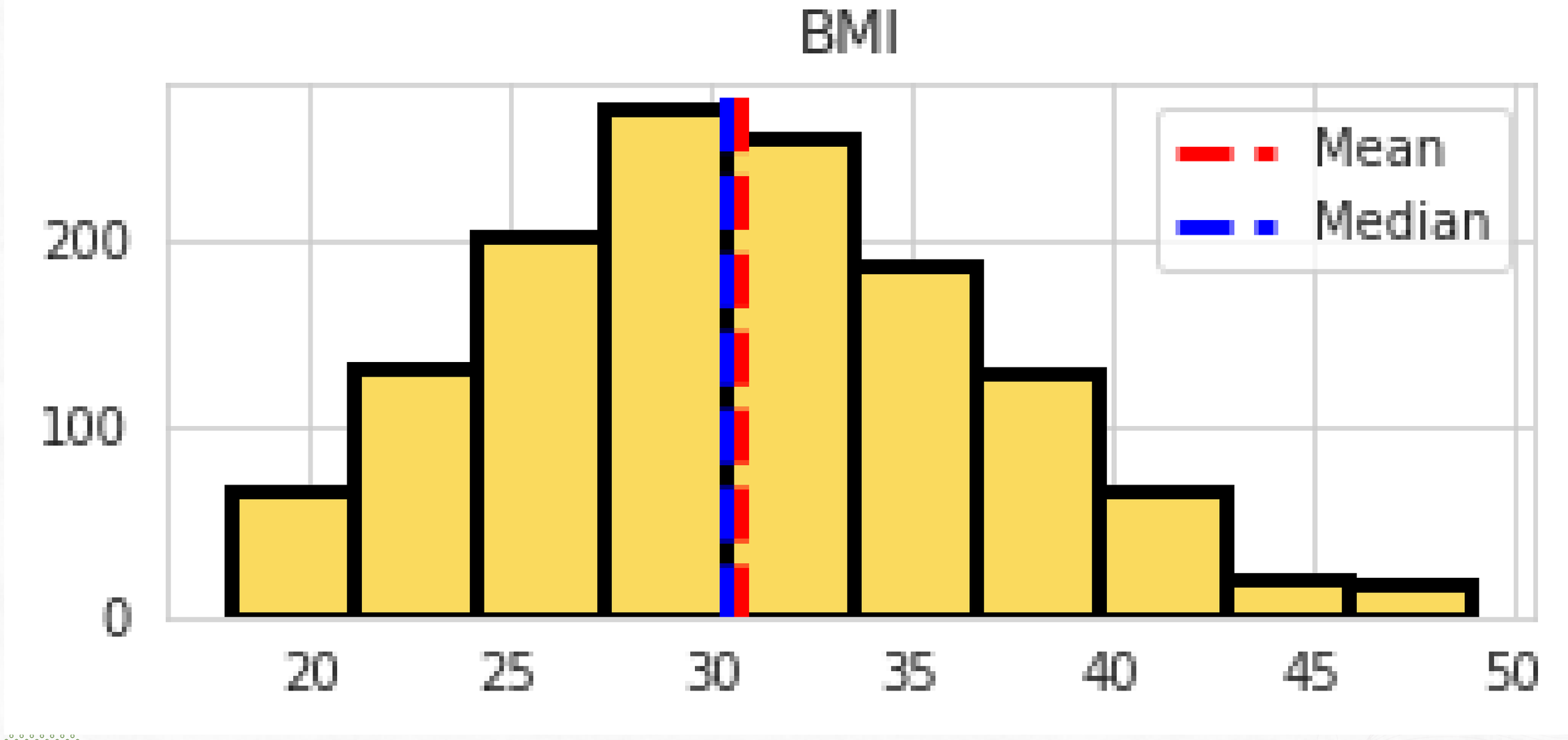

Region Percent
northwest
northeast
24.29%(325)
24.22%(324)
None
27.20%(364)
24.29%(325)
southeast
southwest


Smoker Percent
no
79.52%(1064)
None
20.48%(274)
yes


Sex Percent
female
49.48%(662)
None
50.52%(676)
male

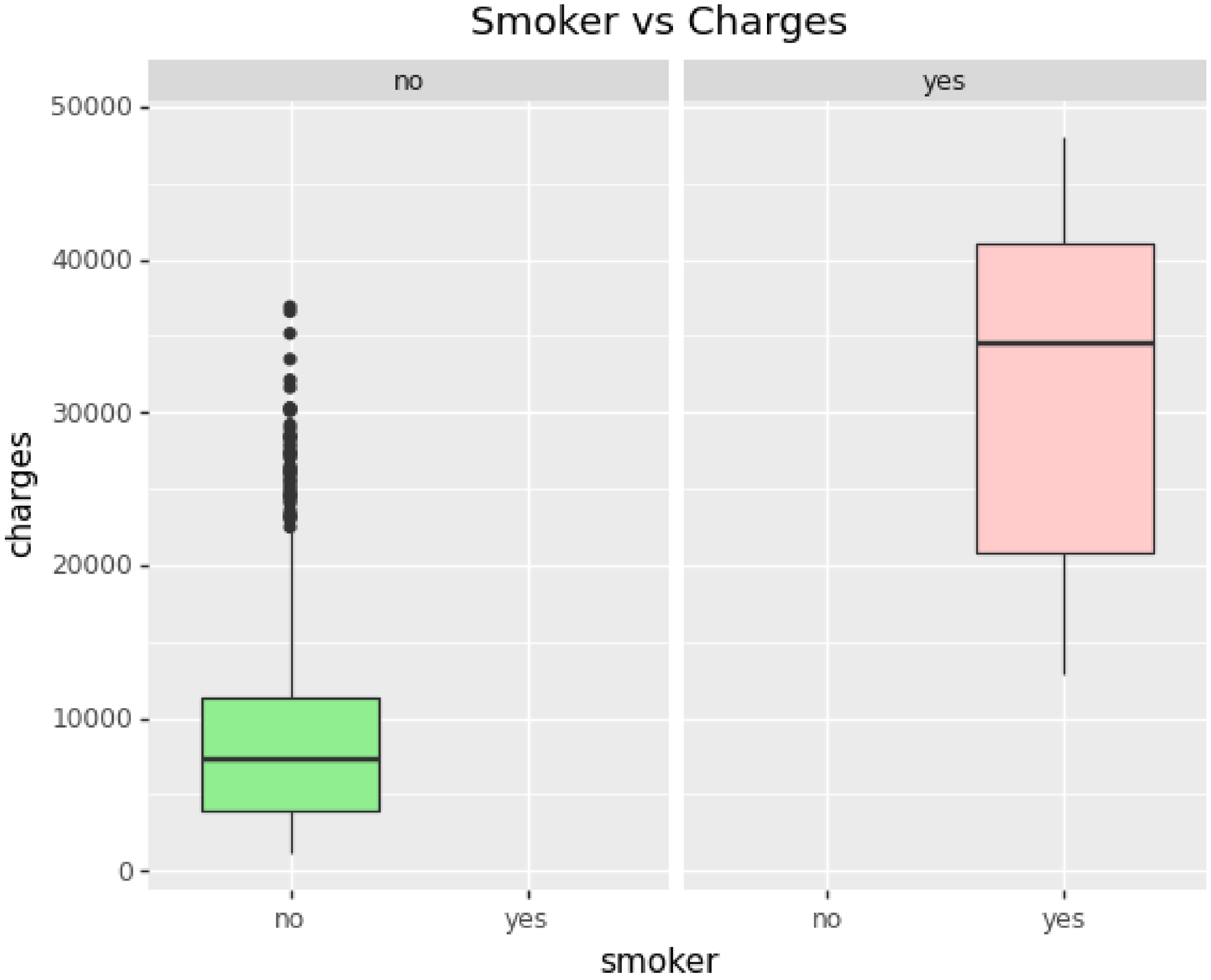

Children Percent
0
42.90%(574)
1
24.22%(324)
2
17.94%(240)
3
11.73%(157)
4
1.87%(25)
5
1.35%(18)
None


Charges
NUMBER OF PEOPLE
400
300
200
100
0
Mean
Median
0
10000
20000
30000
40000
50000
CHARGES


BMI
Mean
Median
200
100
0
20
25
30
35
40
45
50
NUMBER OF PEOPLE
BODY MASS INDEX

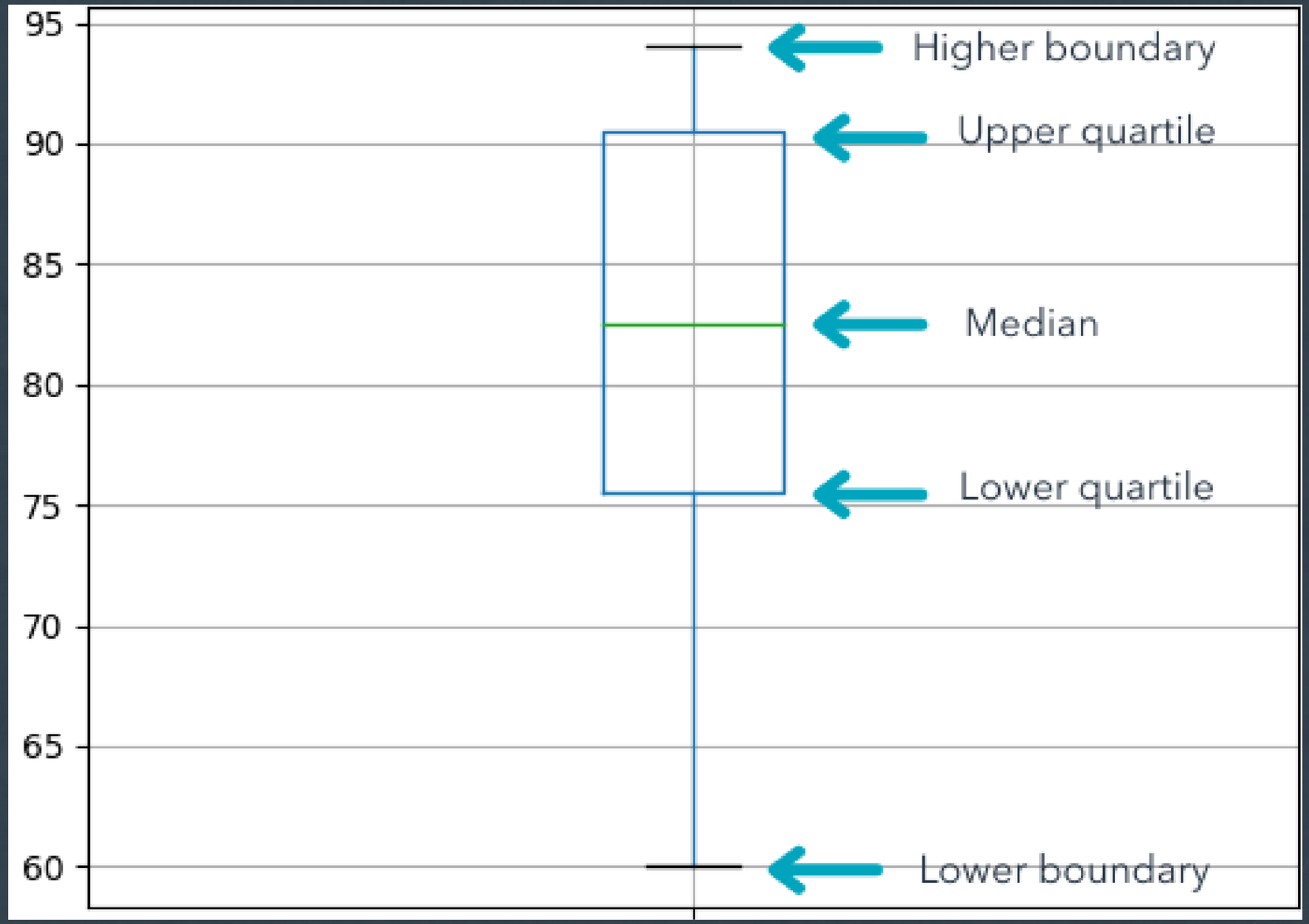
# HISTOGRAM

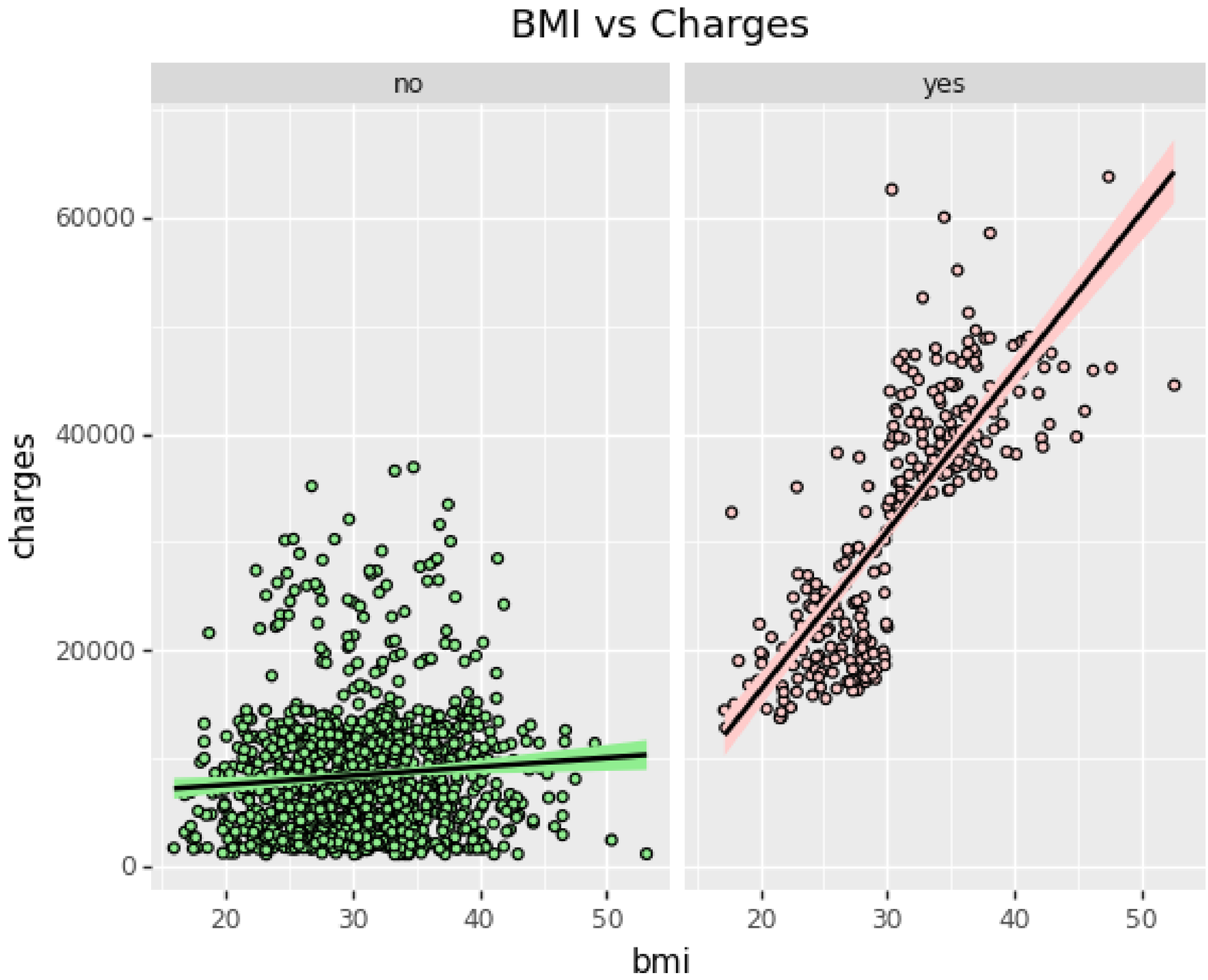
A histogram is a graph showing frequency distributions.
It is a graph showing the number of observations within each given interval.


Smoker vs Charges
no
yes
50000
40000
30000
20000
10000
0
charges
no
yes
no
yes
smoker

# BOXPLOT

Boxplots are a measure of how well distributed the data in a data set is. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile, and third quartile in the data set. It is also useful in comparing the distribution of data across data sets by drawing boxplots for each of them.


95
90
85
80
75
70
65
60
Higher boundary
Upper quartile
Median
Lower quartile
Lower boundary

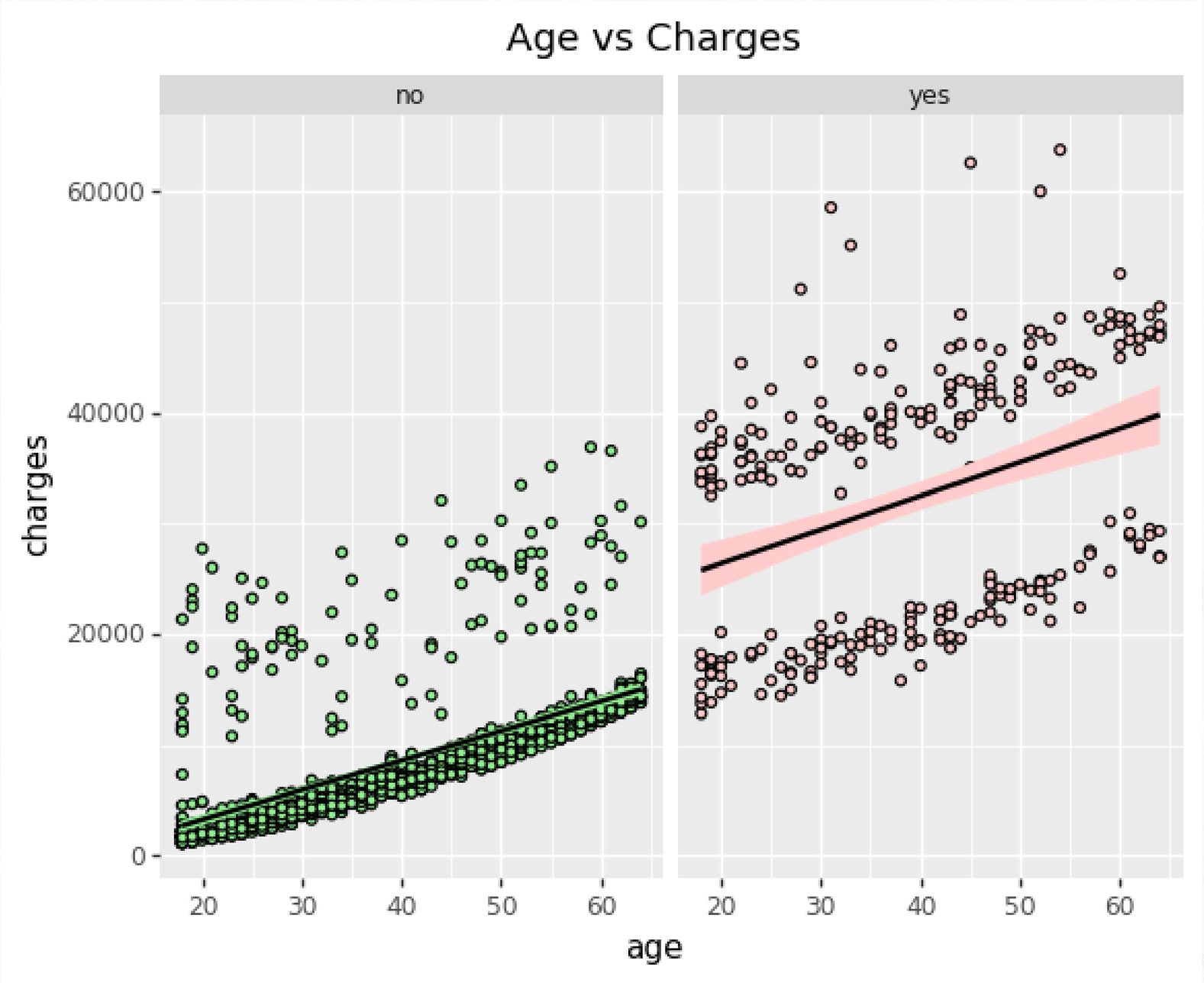



OUTPUTS:

DATA TENDS TO REMAİN STABLE FOR NON-SMOKERS

DATA TENDS TO İNCREASE LİNEARLY FOR SMOKERS.

YES AND NO FOR SMOKERS OR NON-SMOKERS

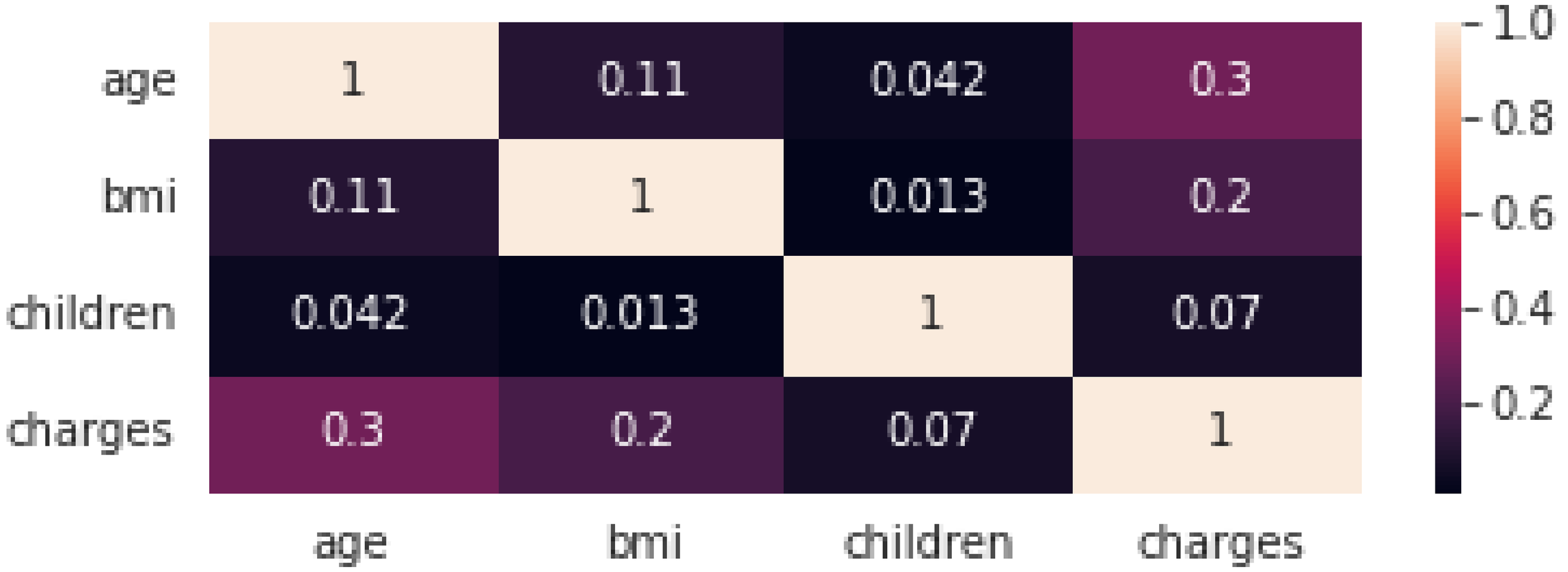



# YES AND NO FOR SMOKERS OR NON-SMOKERS

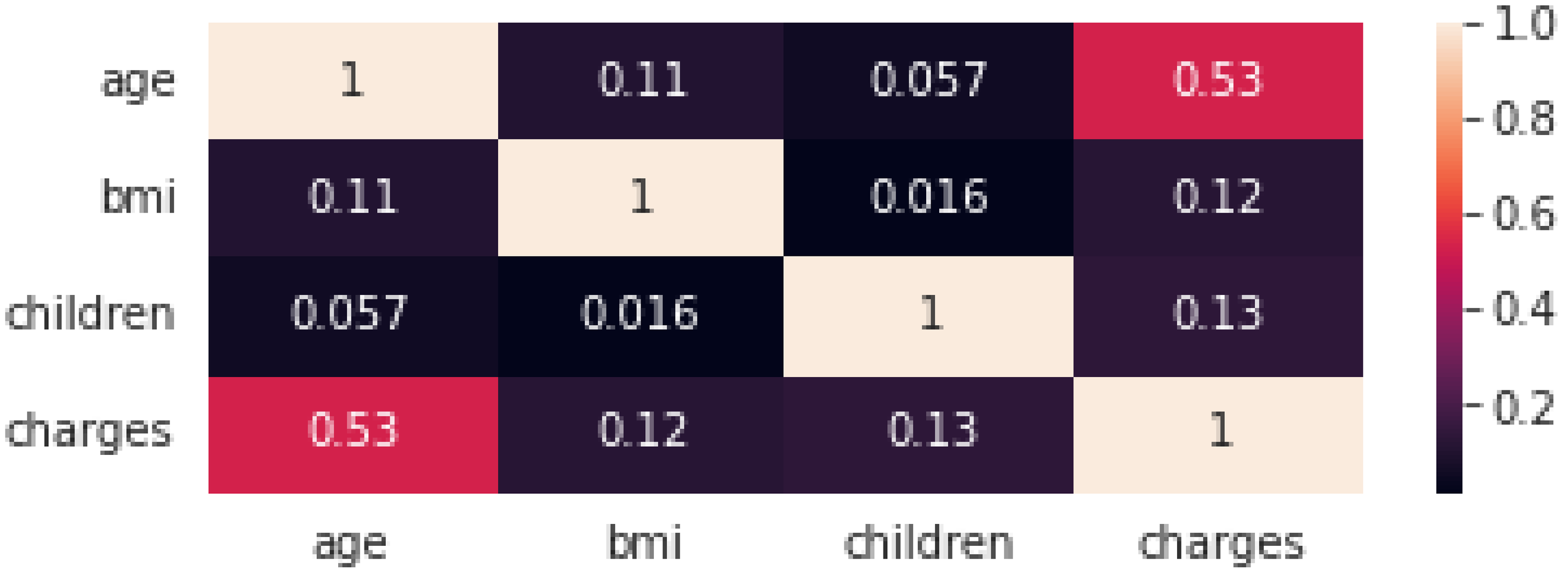
# SCATTER PLOTS


A scatter plot is a diagram where each value in the data set is represented by a dot.

# CORRELATION MATRIX PEARSON



# CORRELATION MATRIX SPEARMAN



# CORRELATION MATRIX

A correlation is a statistical measure of the relationship between two variables. A correlation matrix is a table containing correlation coefficients between variables. Each cell in the table represents the correlation between two variables. The value lies between -1 and 1.

Outputs:

Four clusters are observed.

- Healthy people who do not smoke.
- People who do not smoke but have health problems.
- People who smoke but are healthy.
- People who smoke and have health problems.

Depending on the age of the users or not, their health status and smoking habits affect the insurance fee.

# 2. FEATURE ENGINEERING

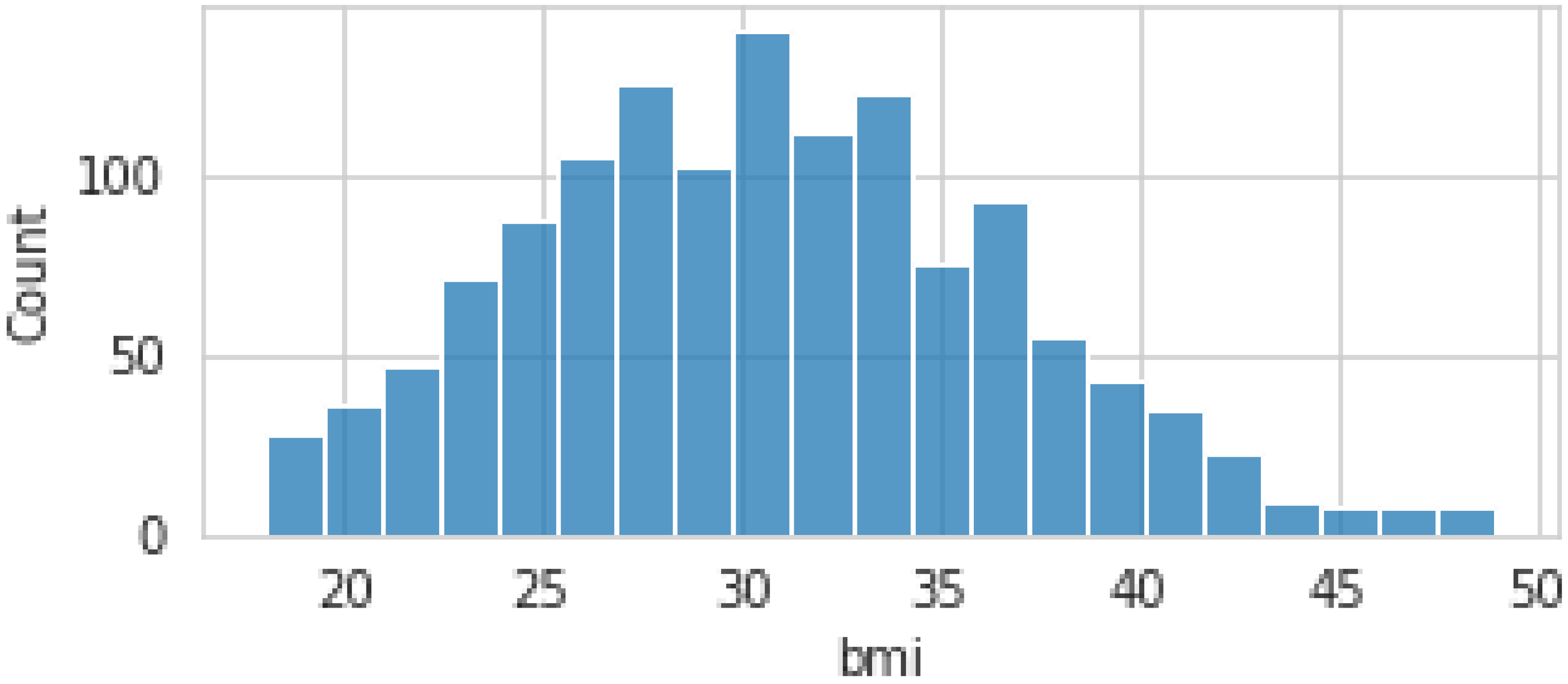
Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.

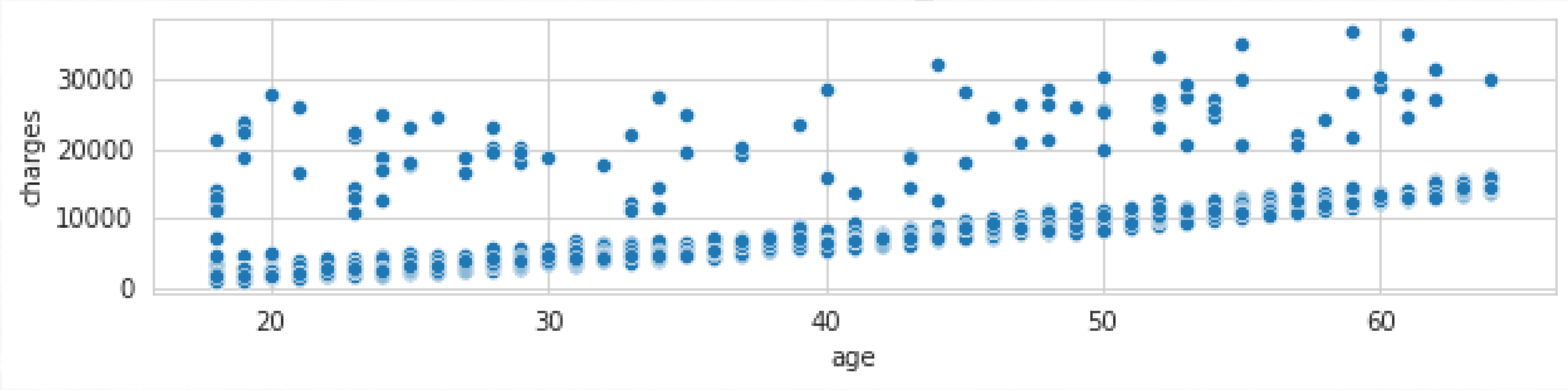
Feature engineering, in simple terms, is the act of converting raw observations into desired features using statistical or machine-learning approaches.

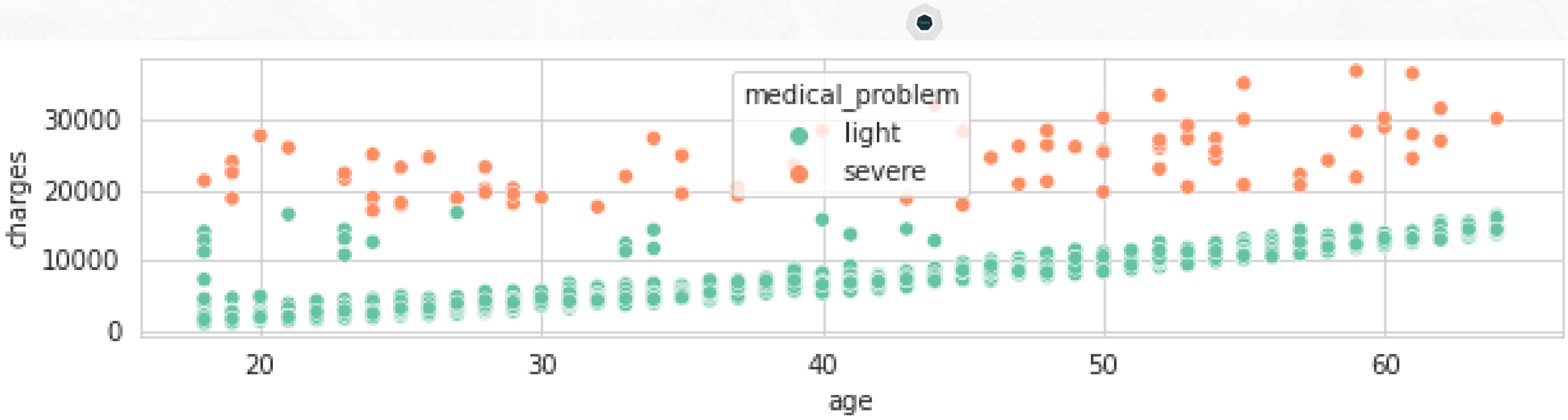
# 2. FEATURE ENGINEERING

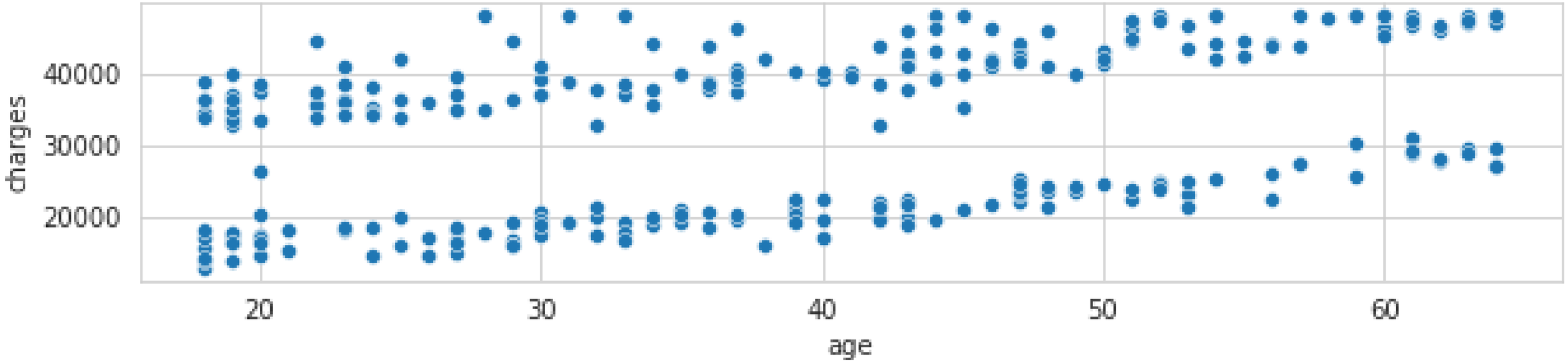
What to do in preparation for this part:
Deleting any null values found and replacing them with logical and appropriate data.
Adding the missing medical problems in the data set.
Intervals are immutable objects that can be created by specifying their connected components. Studies will be made to find the best interval.

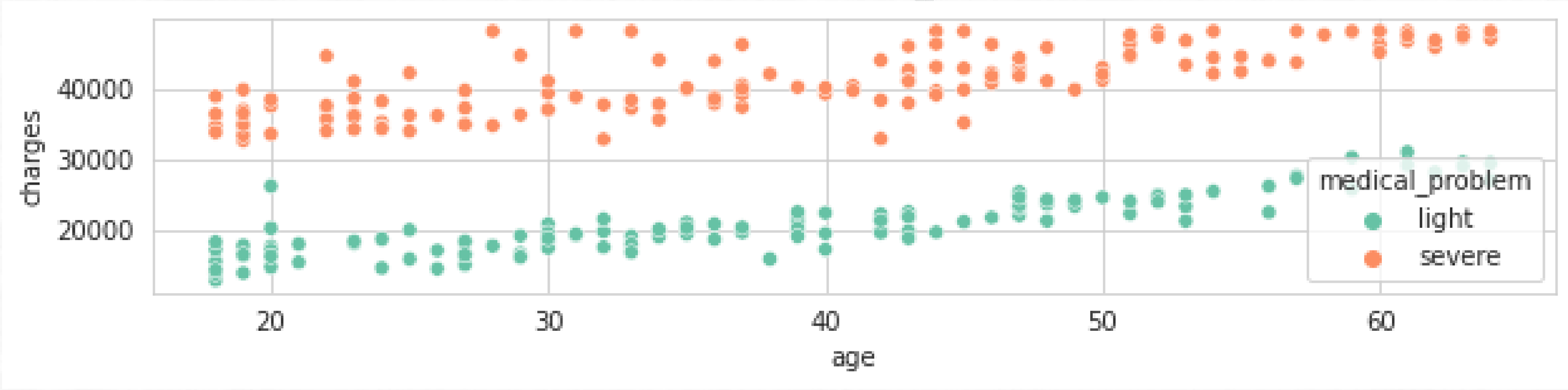
After finding the upper and lower ranges through the code, the author repeats the EDA after determining the reasonable upper and lower limits.

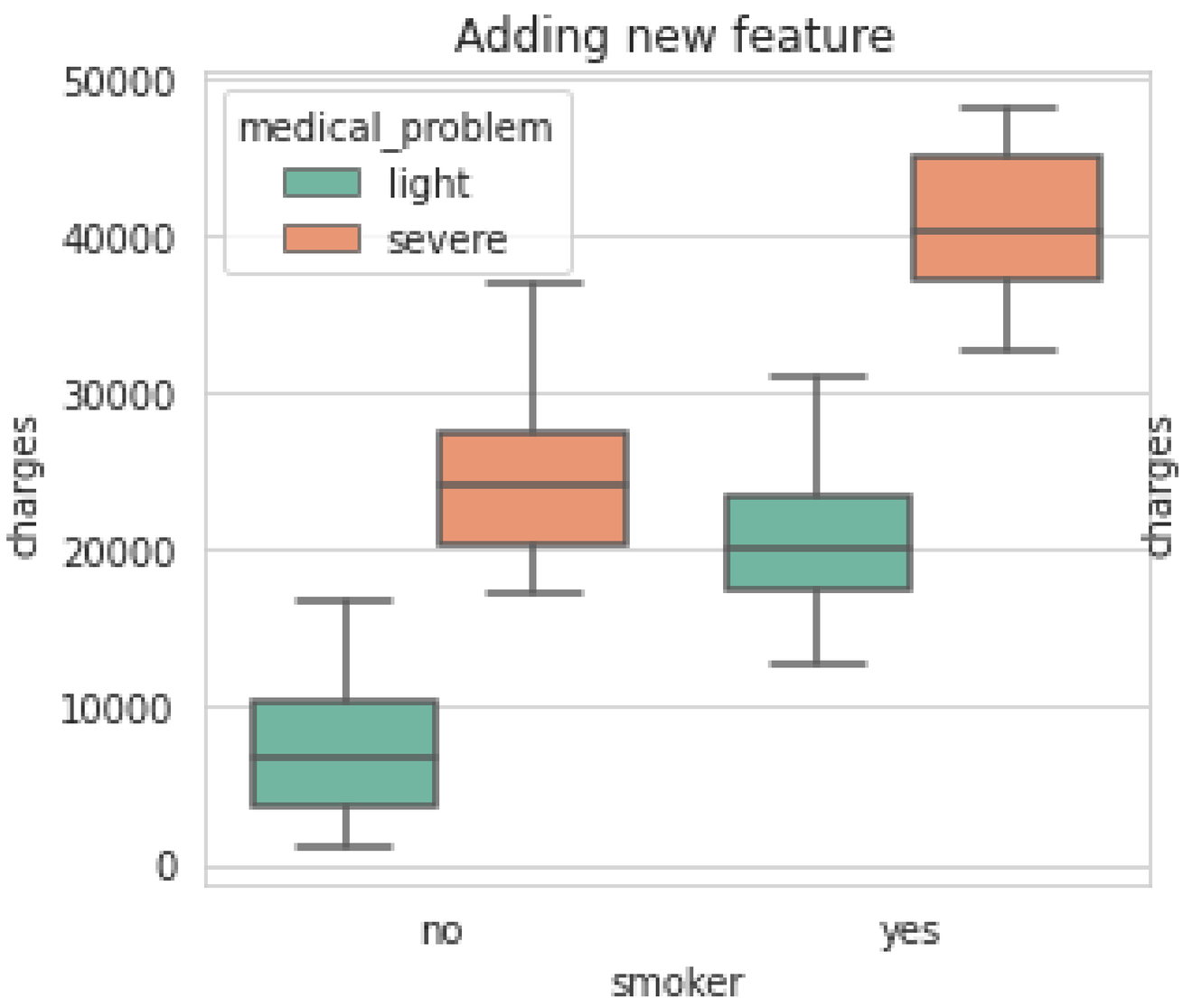

Count
100
50
0
20
25
30
35
40
45
50
bmi

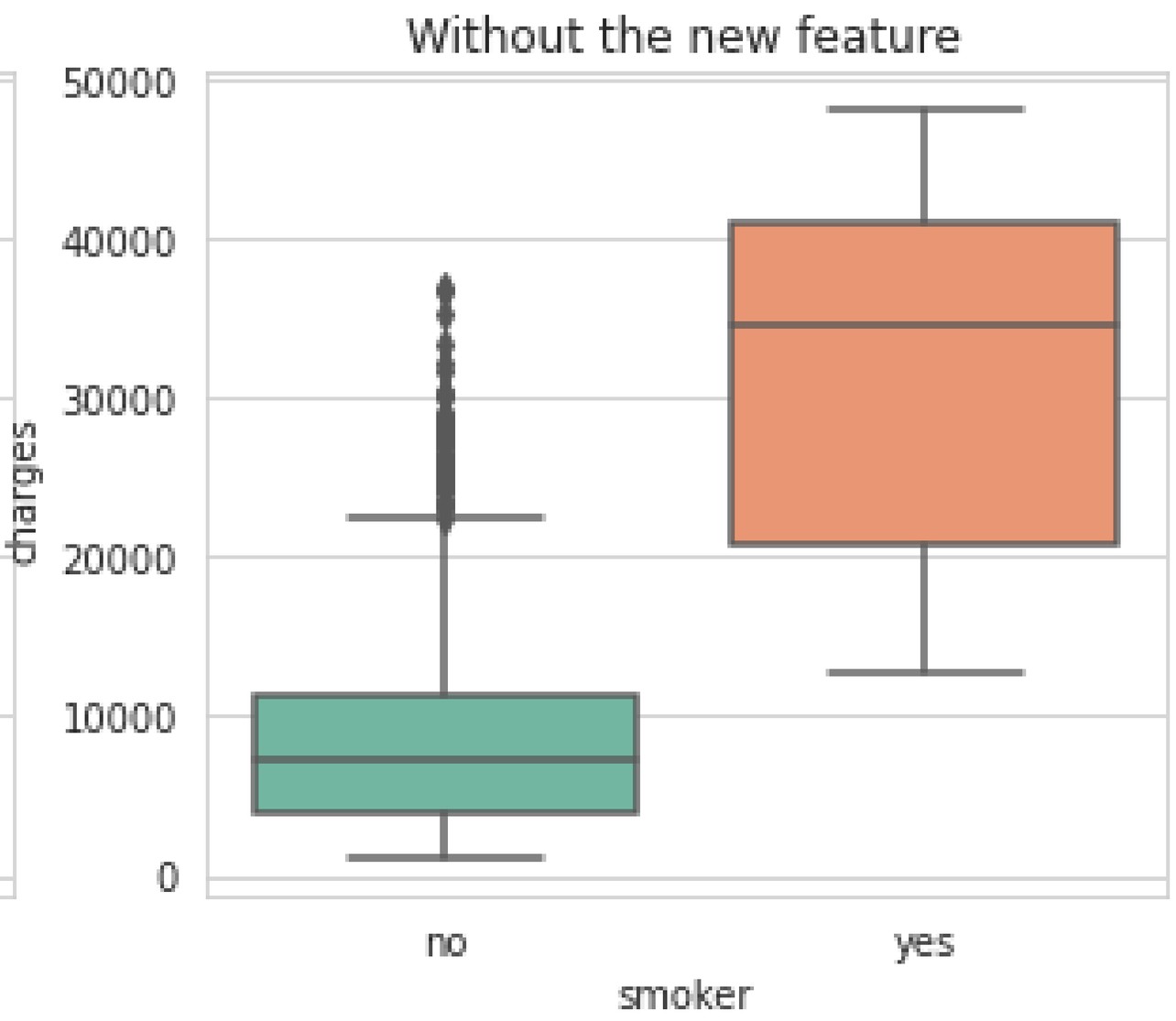

charges
30000
20000
10000
0
20
30
40
50
60
age


medical_problem
light
severe
charges
30000
20000
10000
0
20
30
40
50
60
age


charges
40000
30000
20000
20
30
40
50
60
age


charges
40000
30000
20000
20
30
40
50
60
age
medical_problem
light
severe


Adding new feature
medical_problem
light
severe
50000
40000
30000
20000
10000
0
charges
no
yes
smoker


Without the new feature
50000
40000
30000
20000
10000
0
charges
no
yes
smoker

# 3. SELECTION OF MODEL IDEAL

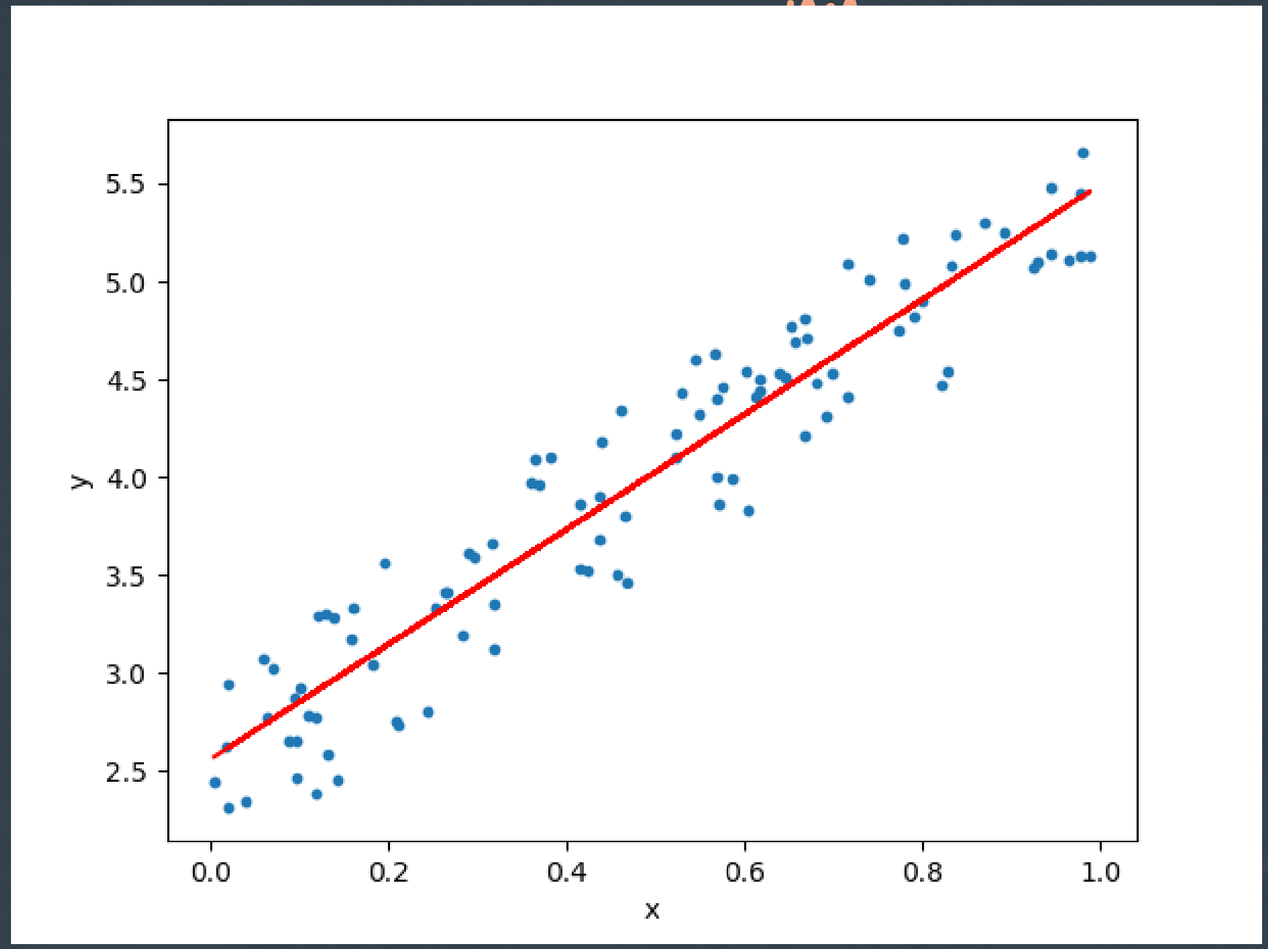
Considering the purpose of the project, it is the part where the models to be used for problem-solving will be tested and the most suitable model for the problem will be selected.

# MODELS TO BE USED:

- Linear Regression
- Gradient Boosting
- XGBOOST

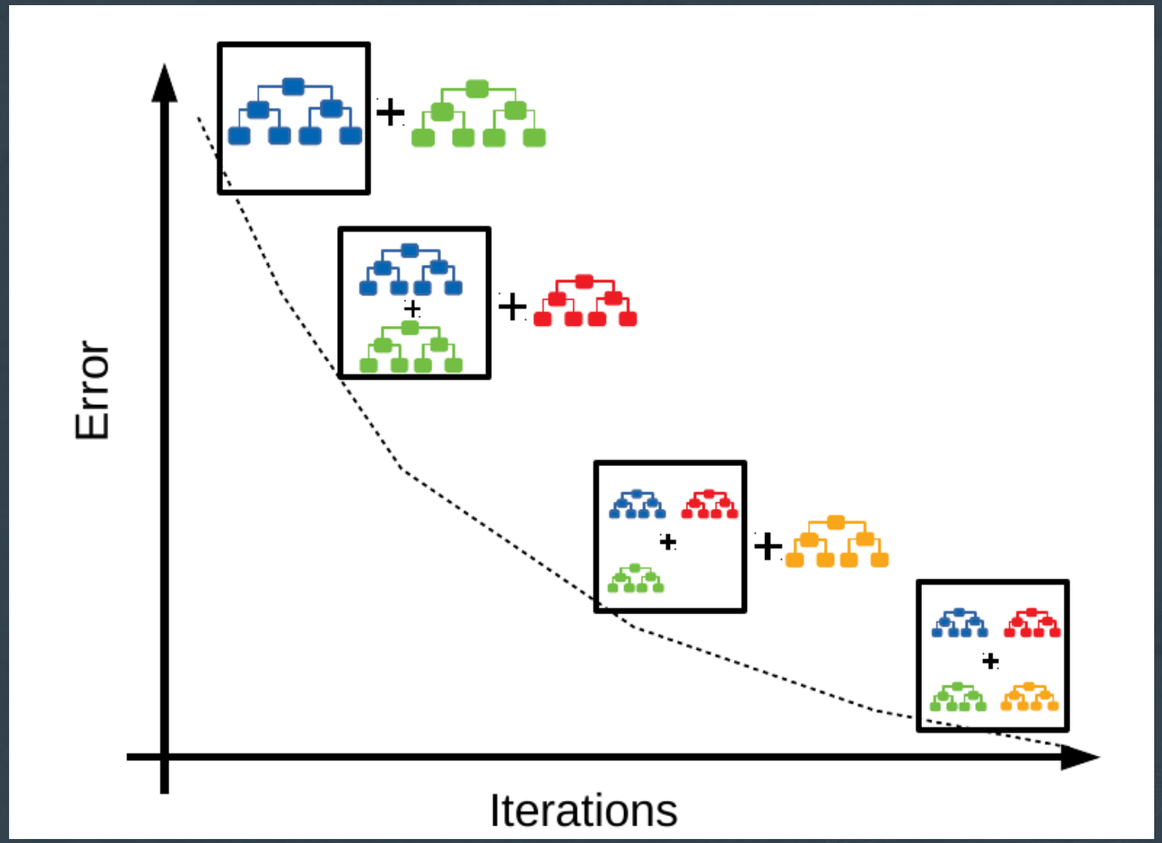
# LINEAR REGRESSION

Linear regression uses the relationship between the data points to draw a straight line through all of them. This line can be used to predict future values. In Machine Learning, predicting the future is quite used.



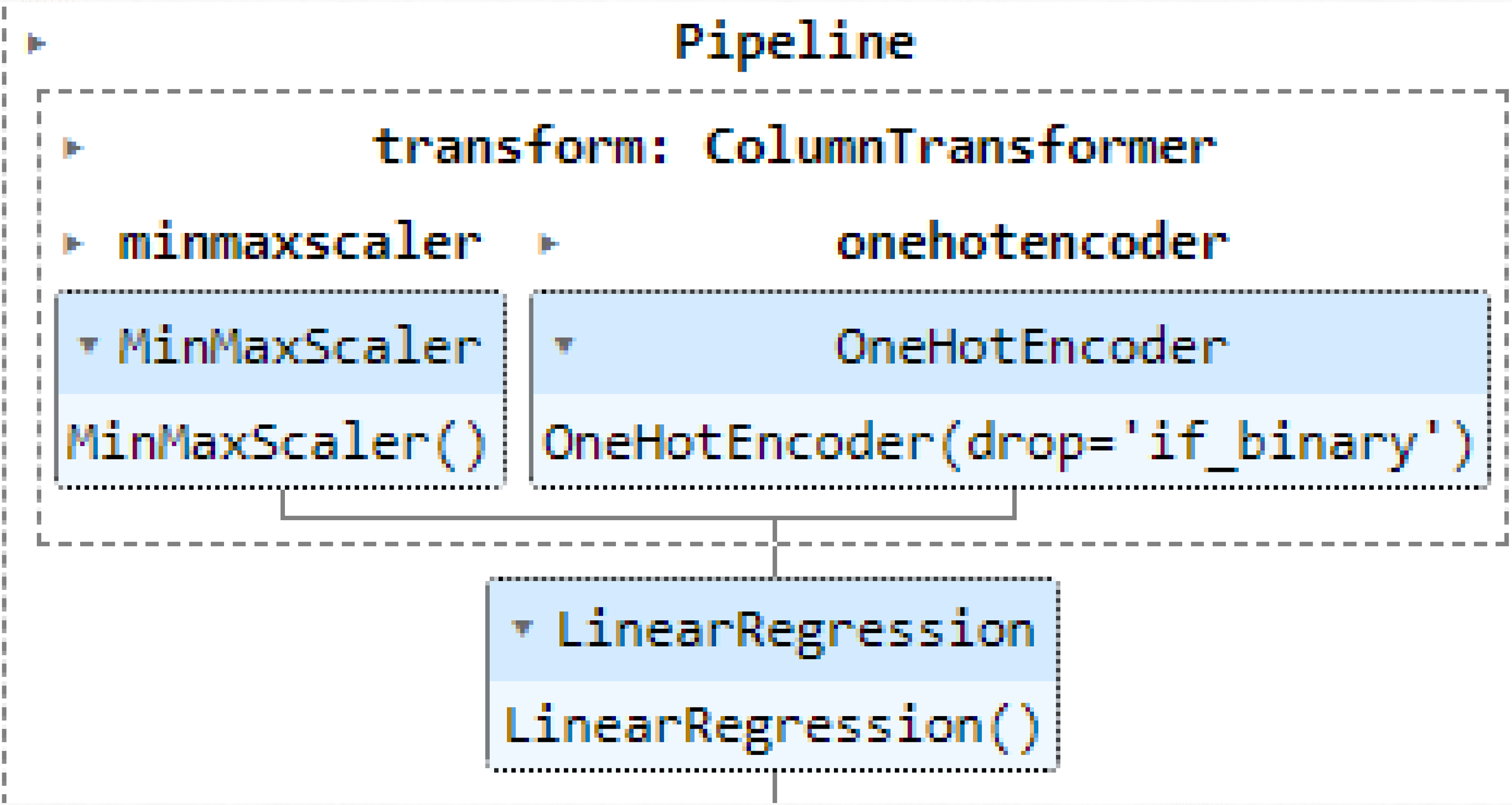
# GRADİENT BOOSTİNG

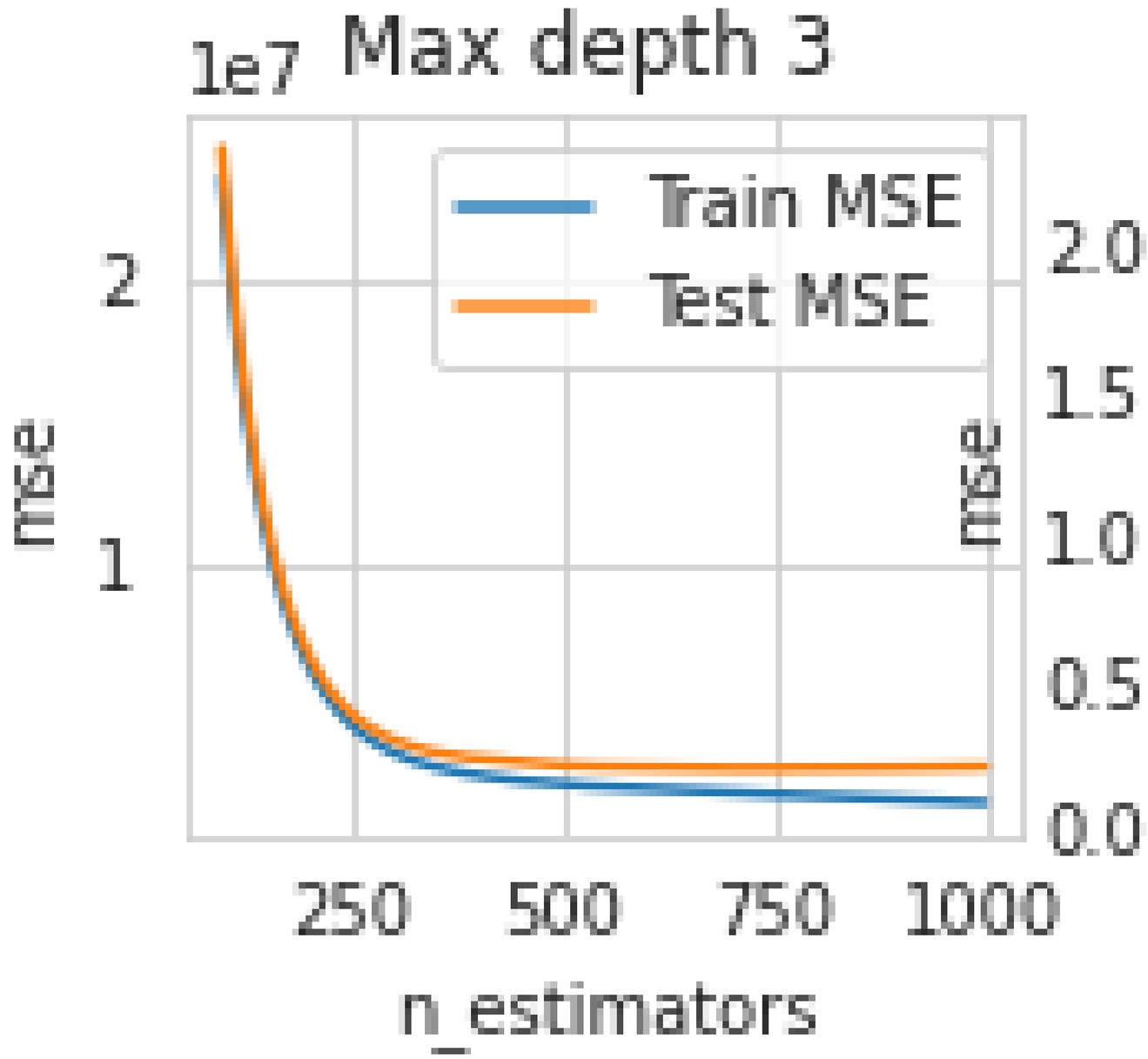
Gradient Boosting is a functional gradient algorithm that repeatedly selects a function that leads in the direction of a weak hypothesis or negative gradient so that it can minimize a loss function.

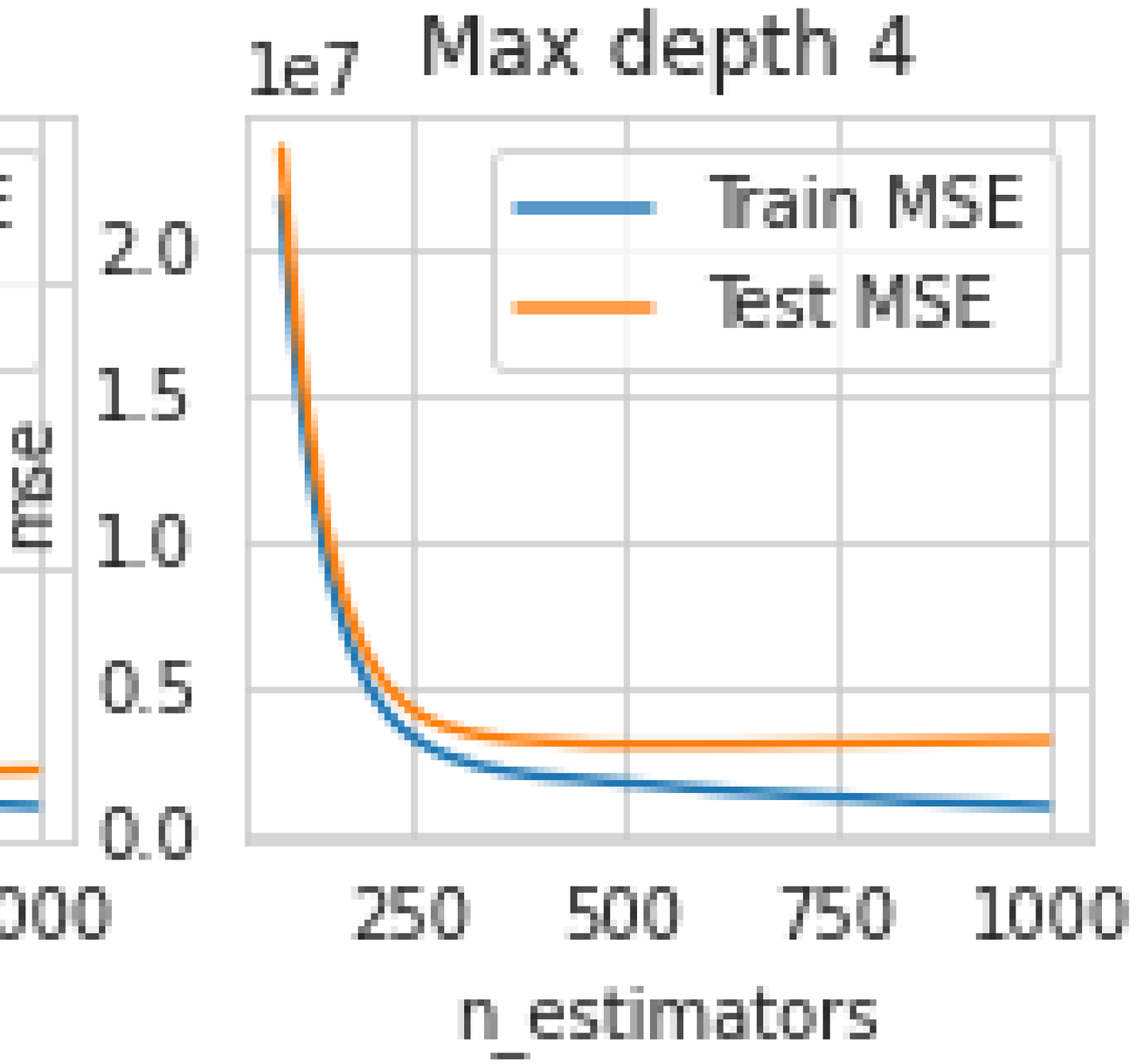


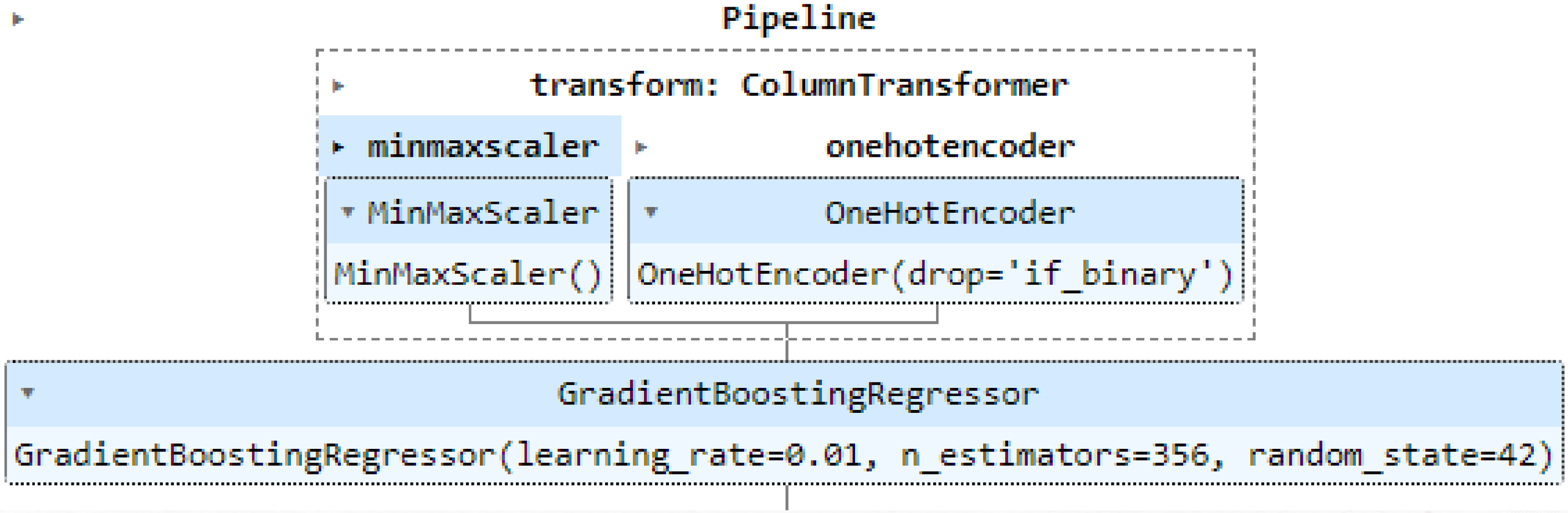
# XGBOOST

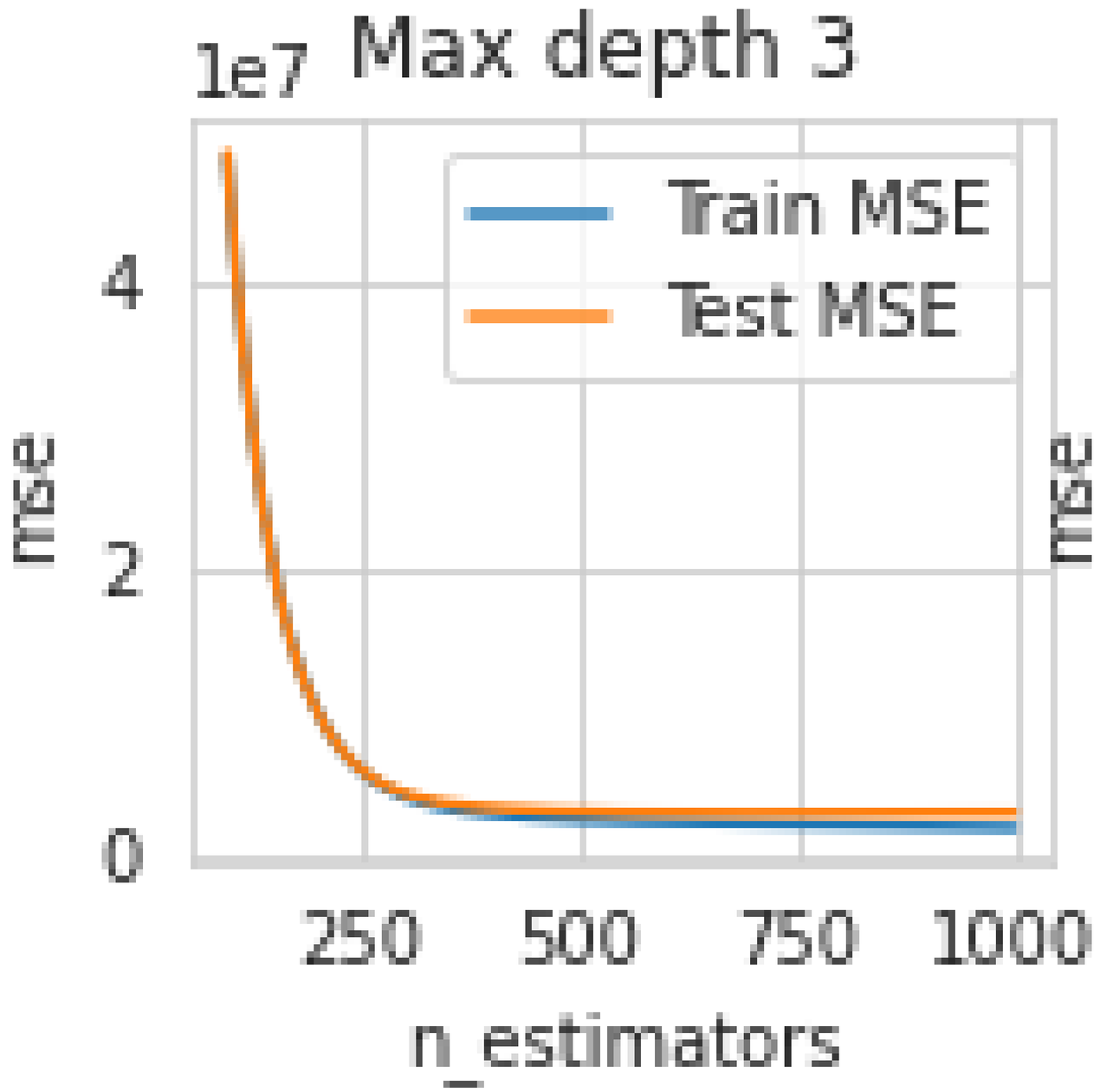
XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine-learning library for regression, classification, and ranking problems.

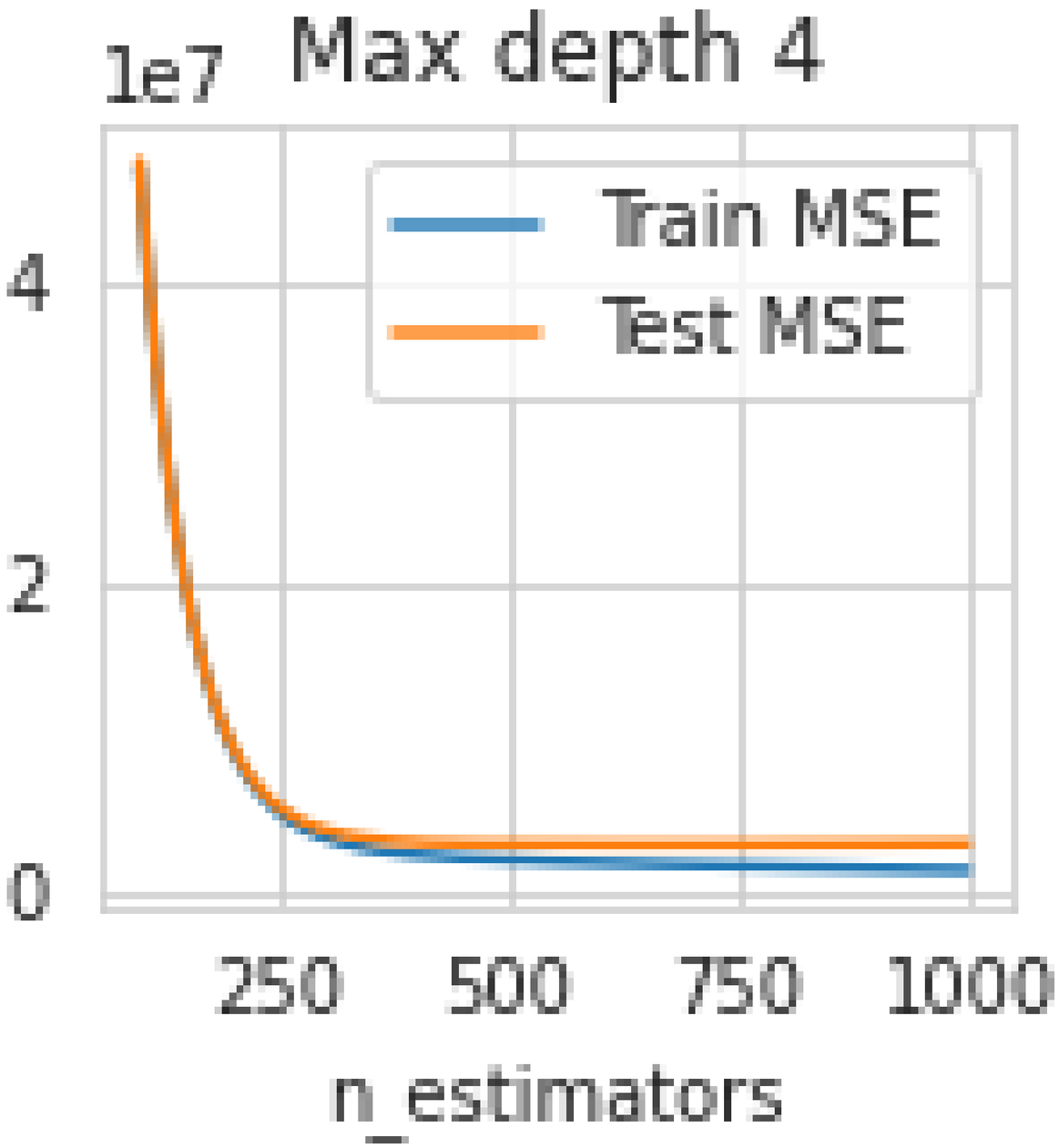

Pipeline
transform: ColumnTransformer
minmaxscaler
onehotencoder
MinMaxScaler
MinMaxScaler()
OneHotEncoder
OneHotEncoder(drop='if_binary')
LinearRegression
LinearRegression()

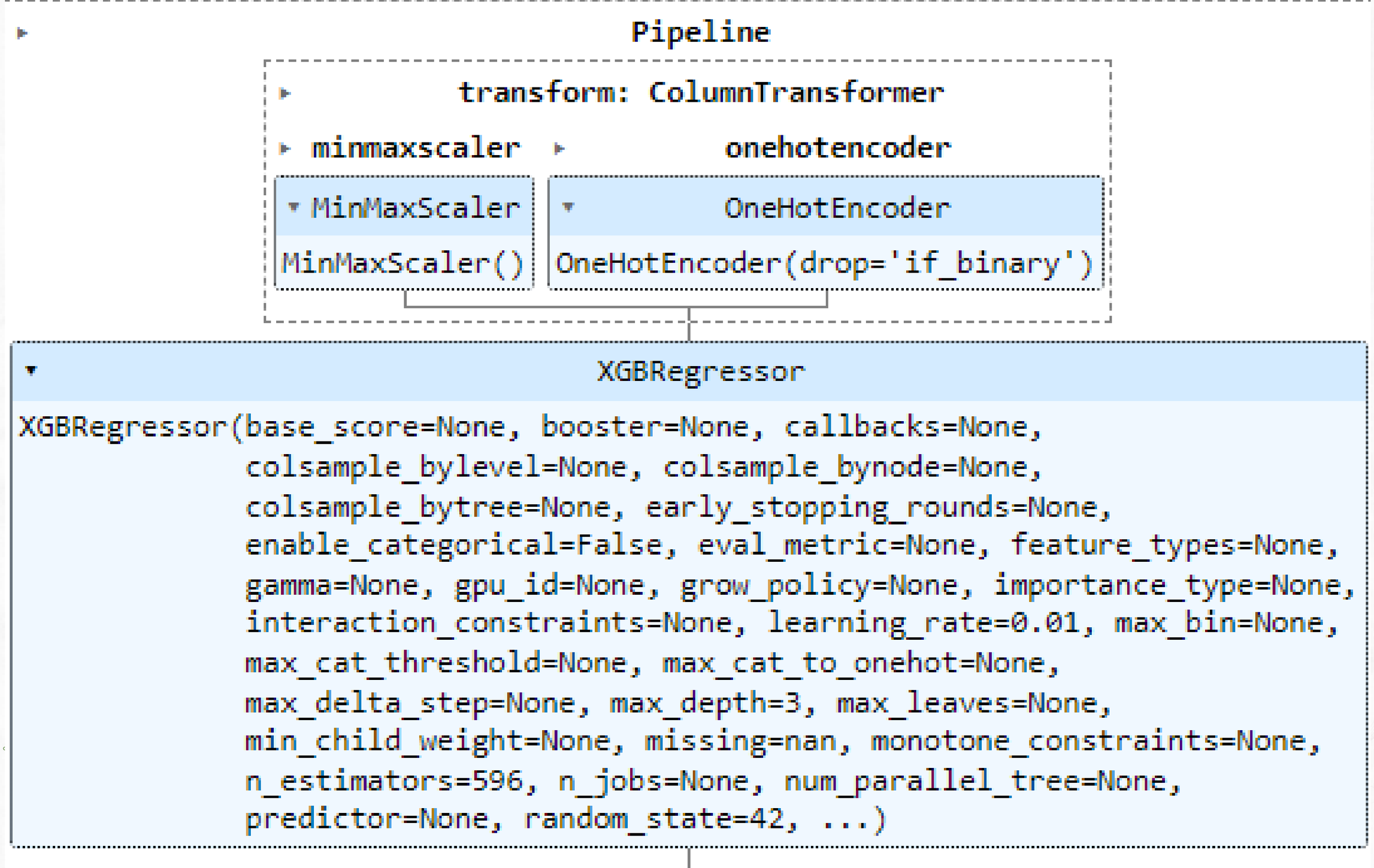

Max depth 3
1e7
Train MSE
Test MSE
mse
2
1
250
500
750
1000
n_estimators


Max depth 4
1e7
Train MSE
Test MSE
mse
2.0
1.5
1.0
0.5
0.0
250
500
750
1000
n_estimators

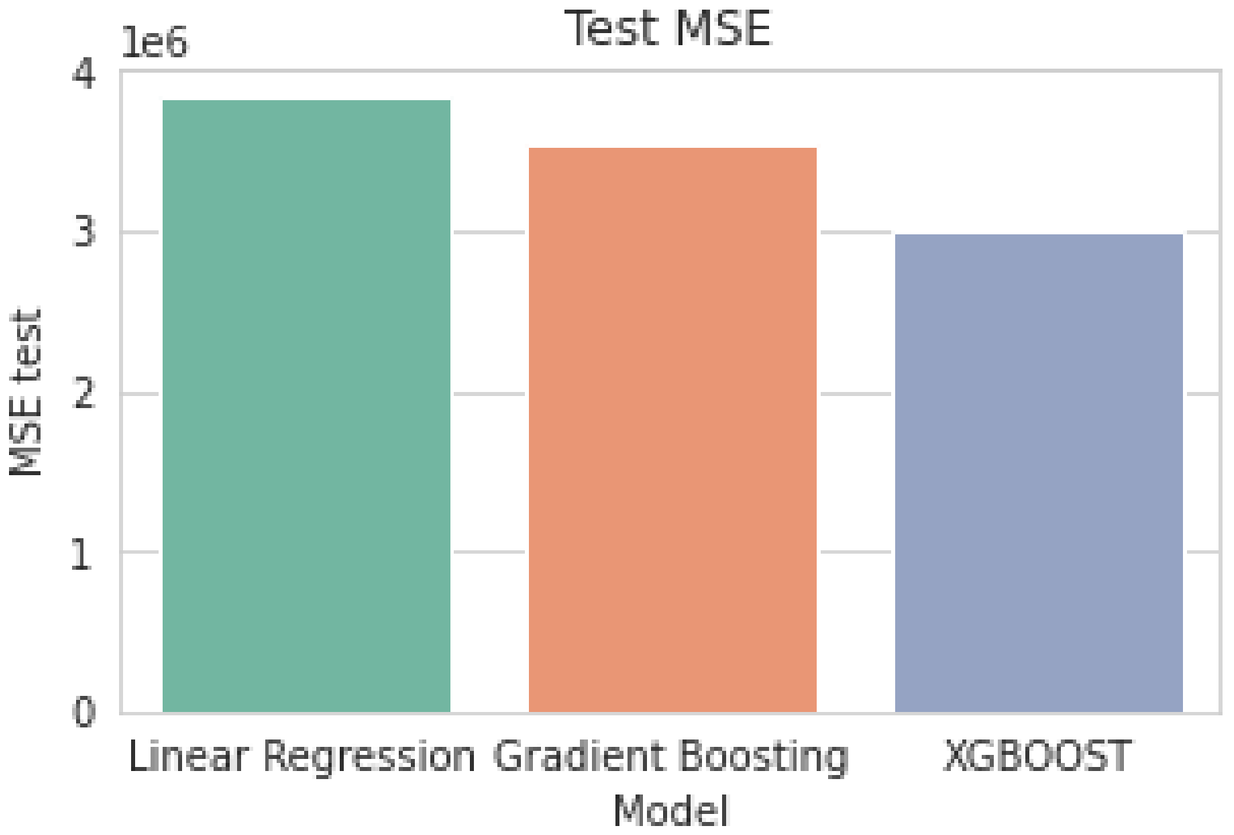

Pipeline
transform: ColumnTransformer
minmaxscaler
onehotencoder
MinMaxScaler
OneHotEncoder
MinMaxScaler()
OneHotEncoder(drop='if_binary')
GradientBoostingRegressor
GradientBoostingRegressor(learning_rate=0.01, n_estimators=356, random_state=42)

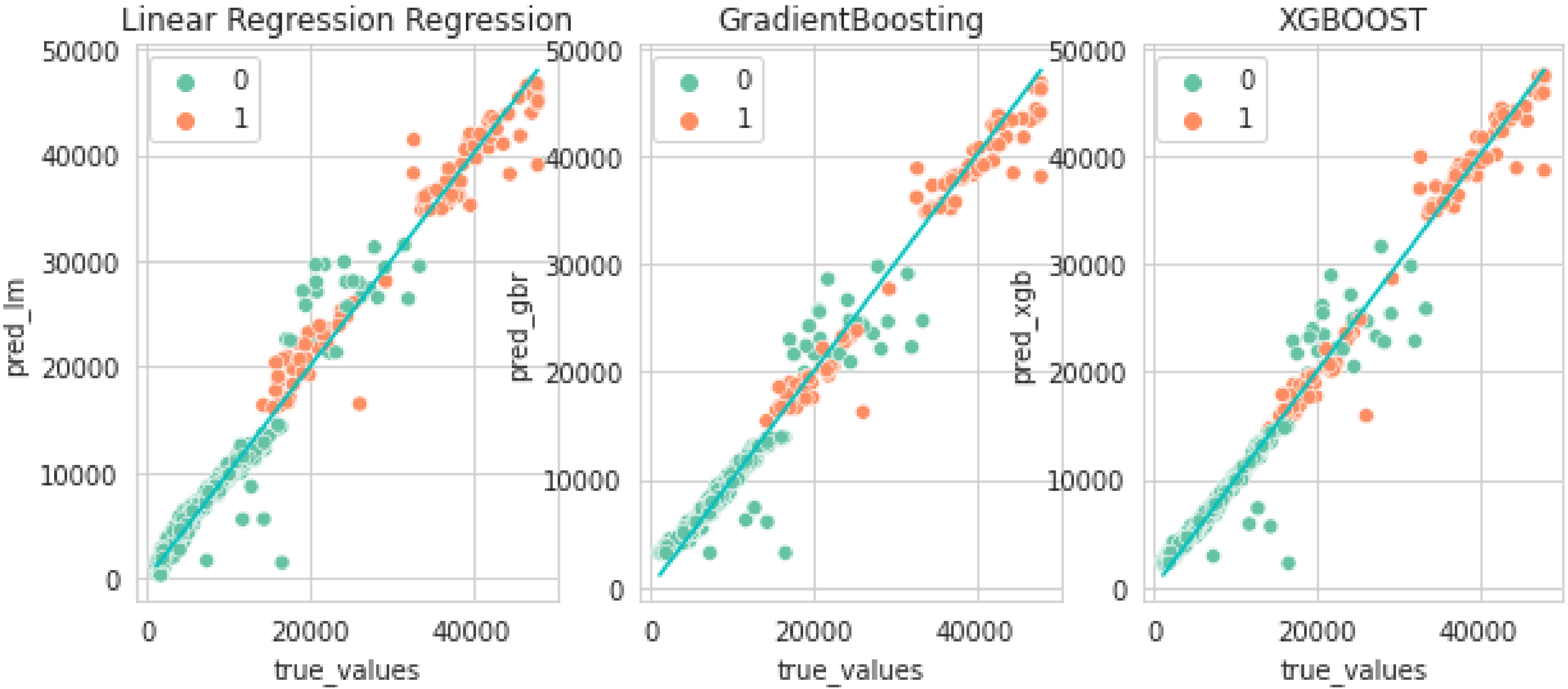

Max depth 3
1e7
Train MSE
Test MSE
mse
4
2
0
250
500
750
1000
n_estimators


Max depth 4
1e7
Train MSE
Test MSE
mse
4
2
0
250
500
750
1000
n_estimators


Pipeline
transform: ColumnTransformer
minmaxscaler
onehotencoder
MinMaxScaler
MinMaxScaler()
OneHotEncoder
OneHotEncoder(drop='if_binary')
XGBRegressor
XGBRegressor(base_score=None, booster=None, callbacks=None,
colsample_bylevel=None, colsample_bynode=None,
colsample_bytree=None, early_stopping_rounds=None,
enable_categorical=False, eval_metric=None, feature_types=None,
gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
interaction_constraints=None, learning_rate=0.01, max_bin=None,
max_cat_threshold=None, max_cat_to_onehot=None,
max_delta_step=None, max_depth=3, max_leaves=None,
min_child_weight=None, missing=nan, monotone_constraints=None,
n_estimators=596, n_jobs=None, num_parallel_tree=None,
predictor=None, random_state=42, ...)


Train MSE
1e6
4
3
2
1
0
MSE train
Linear Regression
Gradient Boosting
XGBOOST
Model

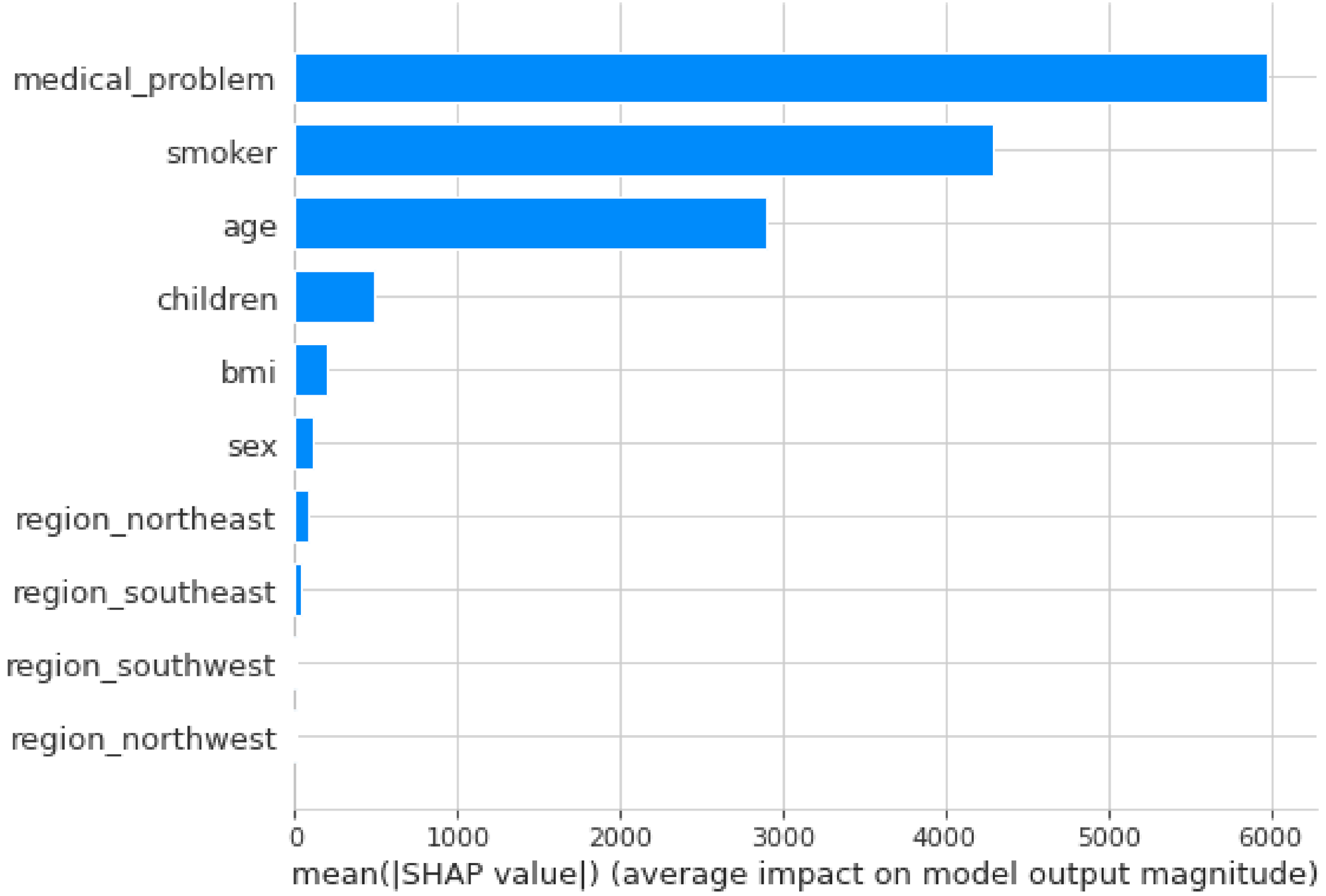

Test MSE
1e6
4
3
2
1
0
MSE test
Linear Regression
Gradient Boosting
XGBOOST
Model

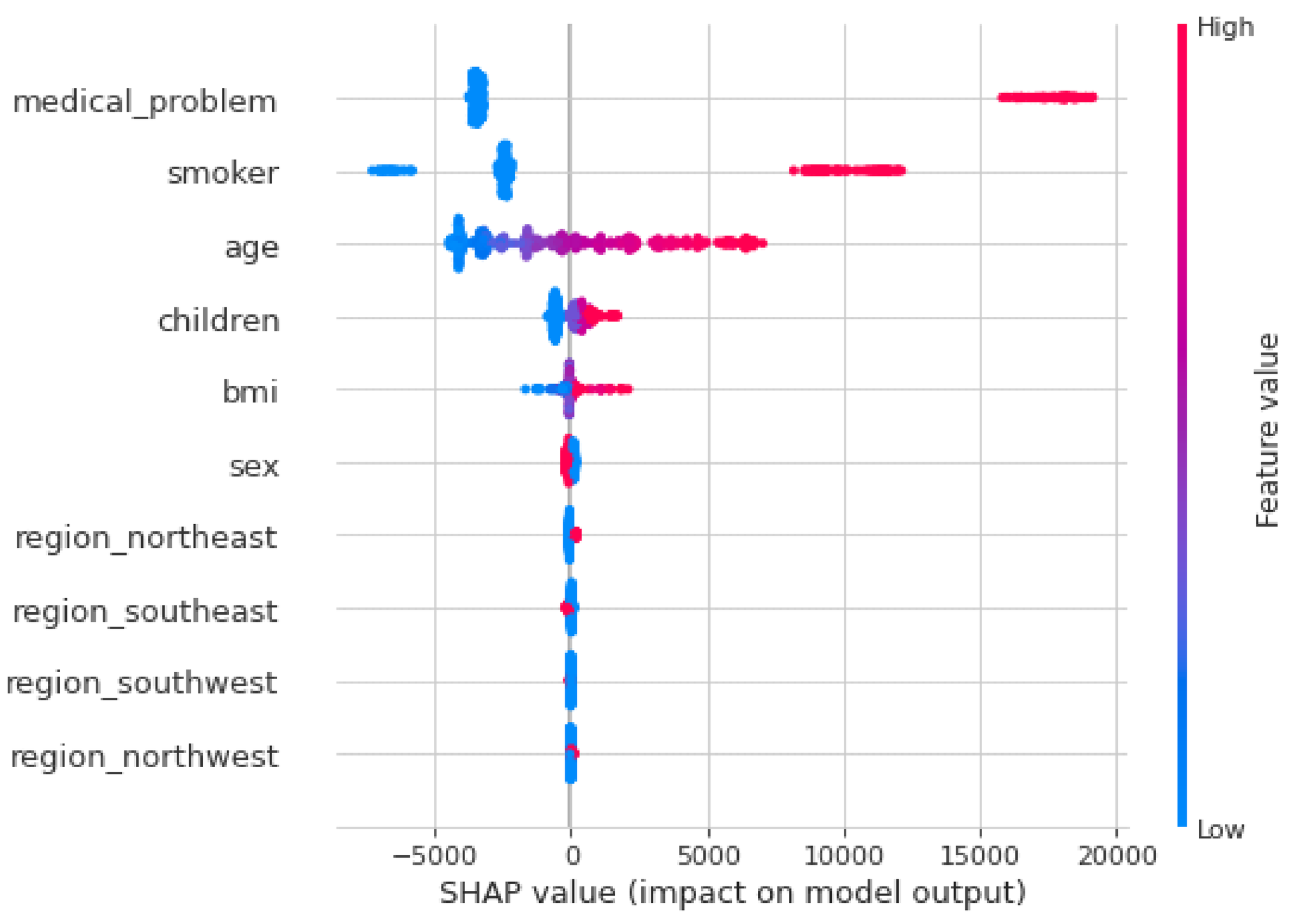

Linear Regression Regression
GradientBoosting
XGBOOST
0
1
pred_lm
pred_gbr
pred_xgb
true_values
0
10000
20000
30000
40000
50000

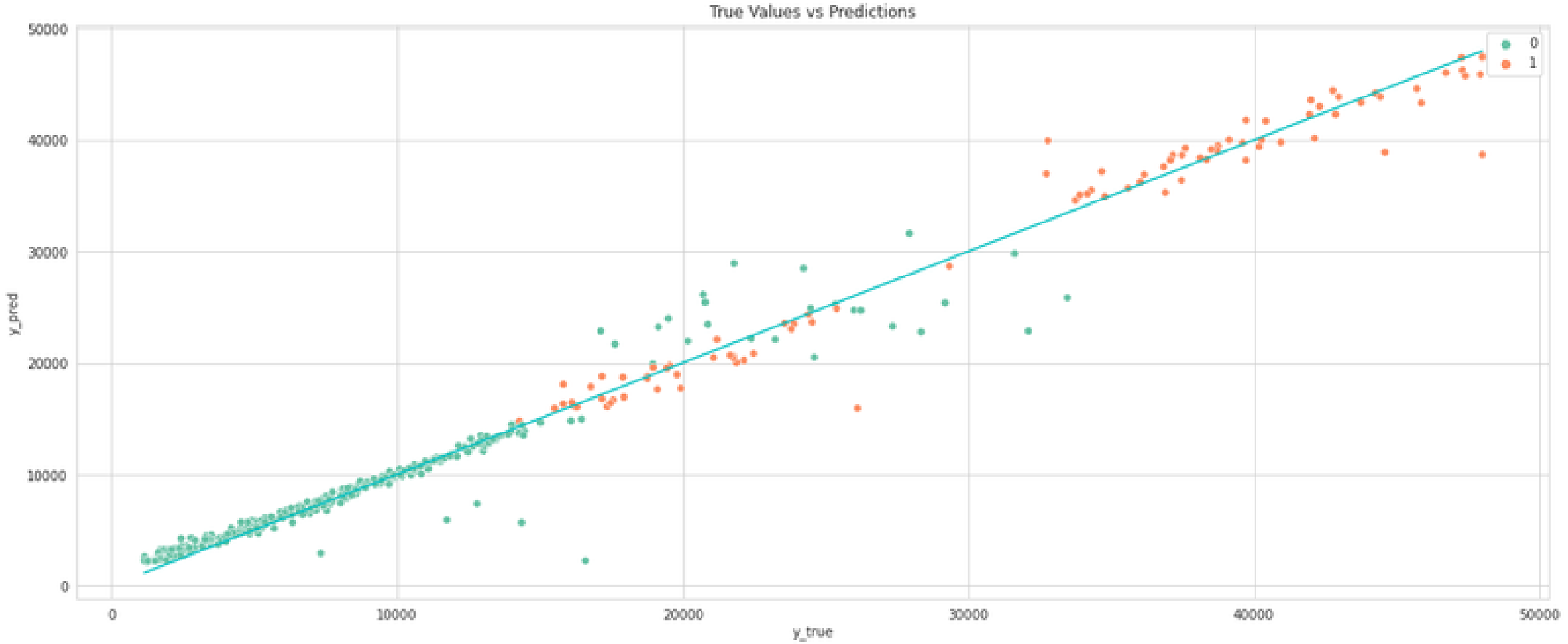
# DEFINITIVE MODEL

▾ XGBRegressor

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.01, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=3, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             n_estimators=596, n_jobs=None, num_parallel_tree=None,
             predictor=None, random_state=42, ...)
```


medical_problem
smoker
age
children
bmi
sex
region_northeast
region_southeast
region_southwest
region_northwest
0
1000
2000
3000
4000
5000
6000
mean(|SHAP value|) (average impact on model output magnitude)


medical_problem
smoker
age
children
bmi
sex
region_northeast
region_southeast
region_southwest
region_northwest
−5000
0
5000
10000
15000
20000
SHAP value (impact on model output)
High
Low
Feature value


True Values vs Predictions
y_pred
y_true
0
10000
20000
30000
40000
50000
0
1

FAUGET CORPORATION

# SALES MARGIN


0,8

## sale margin

Make it your own by customizing it with text and photos.

$1,3b

## revenue

Choose from over a thousand templates to fit any objective or topic.