Estimation and Prediction of Hospitalization and Medical Care Costs

Team ID: LTVIP2023TMID00923

Team Leader: MUDIMADUGU ASWARTHA

Team member: PATHAKANDUKURI GOWTHAMI

Tame member: SHAIK ABDUL RAZAK

Team member: YADDULA GOWTHAMI

INDUSTRY MENTOR: K. INDRA

FACULTY MENTOR: L. RESHMA

FACULTY EVALUATOR:L. RESHMA

OBJECTIVES

- Ensure effective resource allocation and financial planning in health care system.
- Enable informed decisions-making for health care.
- Support cost containment, risk management, and efficient health care delivery.

MEDICAL CHARGES

CONTEXT

Estimating and predicting hospitalization and medical care costs can be done through data analysis and modeling. By analyzing historical data and relevant variables such as patient demographics, medical conditions, treatment procedures, and length of hospital stay, you can build predictive models using statistical techniques or machine learning algorithms.

Keep in mind that accurate predictions depend on the quality and quantity of data available. Additionally, consider factors such as changes in healthcare policies, advancements in medical technology, and shifts in patient behavior that might affect the accuracy of your predictions. If you have specific data or a particular context in mind, I can help you explore more tailored approaches for estimating and predicting hospitalization and medical care costs.

CONTENT

AGE: age of the primary beneficiary.

SEX: insurance contractor gender, female, male.

BMI: body mass index, providing a an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg /m^2) using the ratio of height to weight, ideally 18.5 to 24.9

CHILDREN: number of the children covered by health insurance / number of dependents.

SMOKER: smoking.

REGION: the beneficiarys residencial area in the US, northeast, southeast, Southwest, Northwest.

CHARGES: individual medical costs billed by health insurance.

LINK TO THE SCRIPT

Estimating and predicting hospitalization and medical care costs require access to relevant data and statistical analysis. To proceed, you would need a dataset containing historical information on hospitalizations, medical procedures, and associated costs. Using this data, you can apply various statistical and machine learning techniques to make predictions.

Common approaches for cost estimation and prediction include linear regression, decision trees, random forests, or even more advanced methods like neural networks. These techniques analyze the relationships between variables to make informed predictions about future costs based on past data.

It's important to note that the accuracy of the predictions largely depends on the quality and quantity of the data available. Additionally, healthcare costs can be influenced by various factors such as patient demographics, medical conditions, insurance coverage, and changes in healthcare policies. If you have a specific dataset or question in mind, feel free to share more details, and I can provide further guidance or assistance.

MODULAS USED IN THE SCRIPT

Estimating and predicting hospitalization and medical care costs usually involves using statistical models and machine learning techniques. Popular methods include linear regression, decision trees, and neural networks. Data on patient demographics, medical history, and treatments are typically used as features. To implement this in a script, you'll need to gather relevant data, preprocess it, and split it into training and testing sets. Then, choose a suitable model, train it on the training data, and evaluate its performance on the test set.

```
# Loading system modules
import sys
import time
import os

# Loading common data related modules
import pandas as pd
import numpy as np
from math import sqrt

# Loading modelling algorithms
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import  GradientBoostingRegressor
from sklearn.linear_model import  Ridge
from sklearn.linear_model import Lasso
from sklearn.linear_model import ElasticNet
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.neural_network import MLPRegressor
from sklearn import linear_model

# Loading tools
from sklearn.model_selection import  train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.metrics import f1_score, precision_score, recall_score
from sklearn.feature_selection import SelectKBest
from sklearn.metrics import mean_absolute_error
from sklearn.utils import shuffle
from sklearn.model_selection import GridSearchCV
import missingno as msno

# Loading visualisation modules
from pandas_profiling import ProfileReport
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt

# Configure visualisations
%matplotlib inline

# Ignore warning messages
import warnings
warnings.filterwarnings('ignore')
from sklearn.utils._testing import ignore_warnings
from sklearn.exceptions import ConvergenceWarning

import plotly.graph_objs as go
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
```

DATA SAMPLE

To estimate and predict hospitalization and medical care costs in a data sample, you can use various statistical and machine learning techniques. Here's a general outline of the process:

1. Data Collection: Gather a comprehensive dataset that includes relevant information such as patient demographics, medical history, diagnosis, treatment details, hospitalization duration, and associated medical costs.

2. Data Preprocessing: Clean the data by handling missing values, outliers, and formatting issues. Ensure that the data is in a suitable format for analysis.

3. Feature Selection: Identify the most relevant features that may impact hospitalization and medical costs, such as age, gender, medical condition, and treatment type.

4. Exploratory Data Analysis (EDA): Perform data visualization and descriptive statistics to gain insights into the relationships between variables and understand the distribution of costs.

5. Model Selection: Choose appropriate machine learning algorithms for prediction. Common choices include linear regression, decision trees, random forests, or gradient boosting.

6. Data Splitting: Divide the dataset into training and testing sets to evaluate the model's performance accurately.

7. Model Training: Train the selected machine learning model using the training data.

8. Model Evaluation: Evaluate the model's performance using the testing data. Common evaluation metrics include Mean Absolute Error (MAE) and Mean Squared Error (MSE).

9. Predictions: Use the trained model to predict hospitalization and medical care costs for new data points.

10. Interpretation: Analyze the model's output and understand the factors that contribute to higher or lower costs.

Remember that the quality of your predictions depends on the quality and representativeness of the data you collected. Also, make sure to follow ethical guidelines regarding data privacy and handling sensitive information when dealing with healthcare data.

	age	sex	bmi	children	smoker	region	charges
0							
1							
2							
3							
4							

```
<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   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 286.5 KB
```

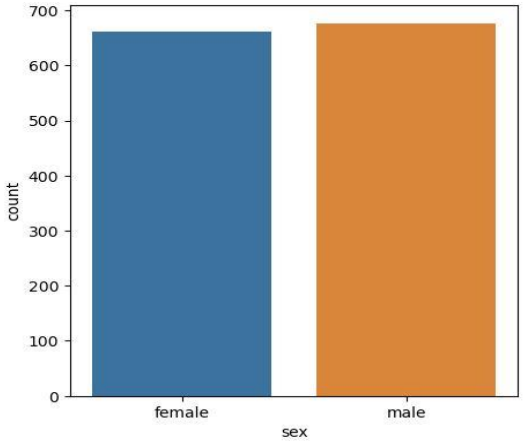
DATA INFORMATION

DATA DESCRIBE

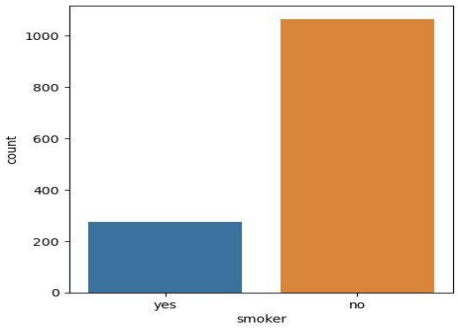
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

CATEGORICAL DATA REVIEW

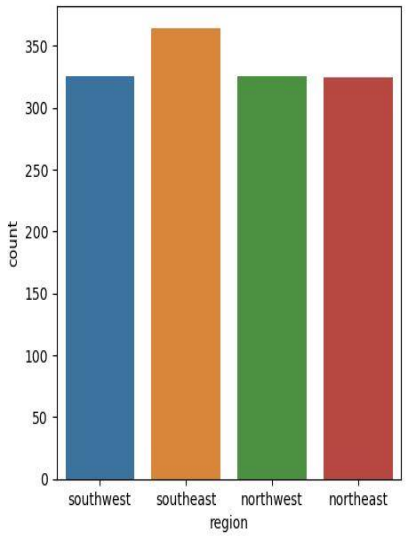
- **SEX**



- **SMOKER**



- **REGION**



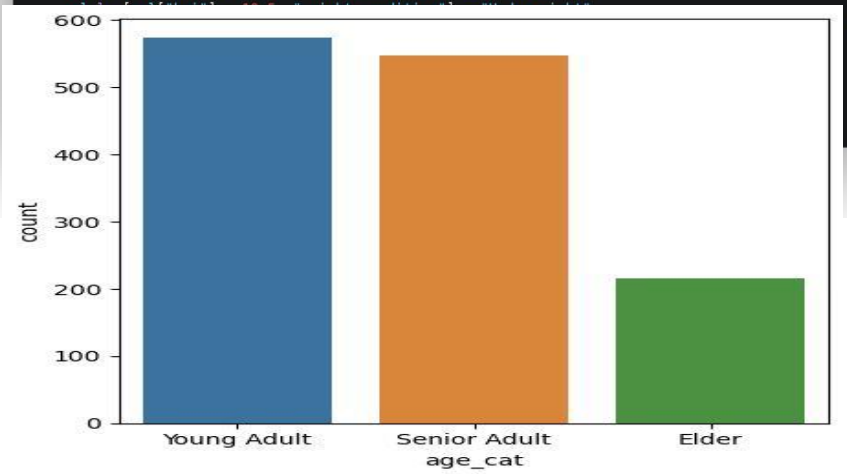
I decided to group the age data and the BMI data into the categories . I suspected that data distribution in this data's is different in each range.

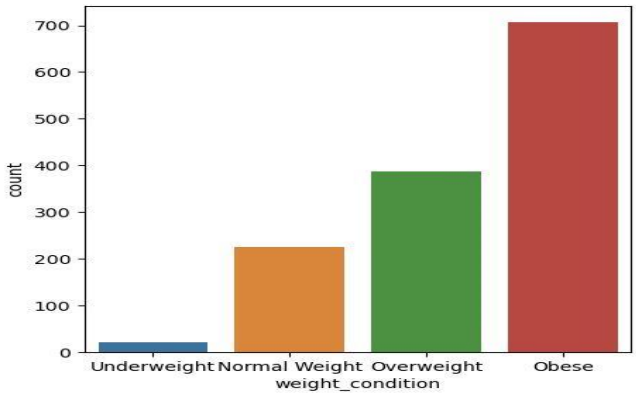
- **AGE**

```
med_cost['age_cat'] = np.nan
lst = [med_cost]

for col in lst:
    col.loc[(col['age'] >= 18) & (col['age'] <= 35), 'age_cat'] = 'Young Adult'
    col.loc[(col['age'] > 35) & (col['age'] <= 55), 'age_cat'] = 'Senior Adult'
    col.loc[col['age'] > 55, 'age_cat'] = 'Elder'
```

- **BMI**

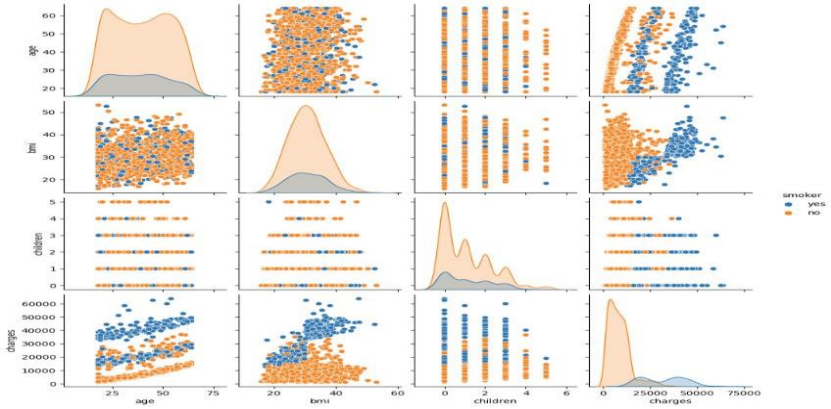
```
med_cost["weight_condition"] = np.nan
lst = [med_cost]

for col in lst:
```



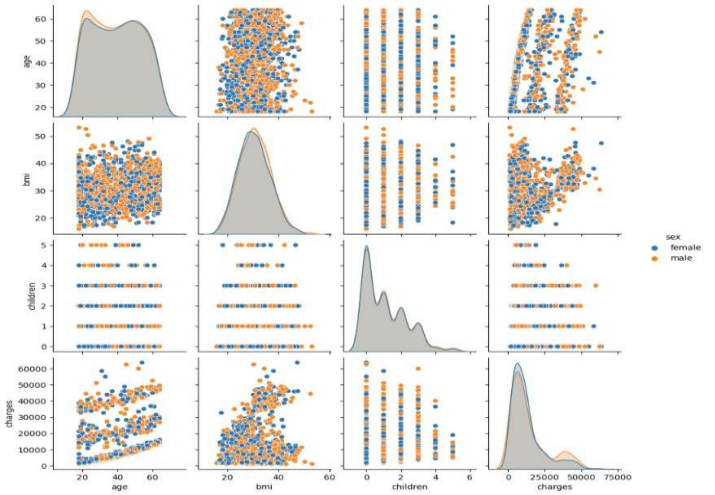


LET'S SEE SOME PAIR PLOT CHARTS

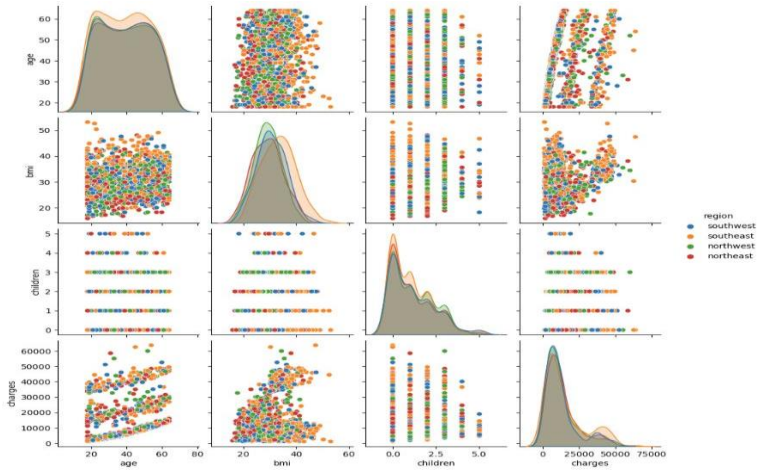
- **SHADE BY SMOKER**



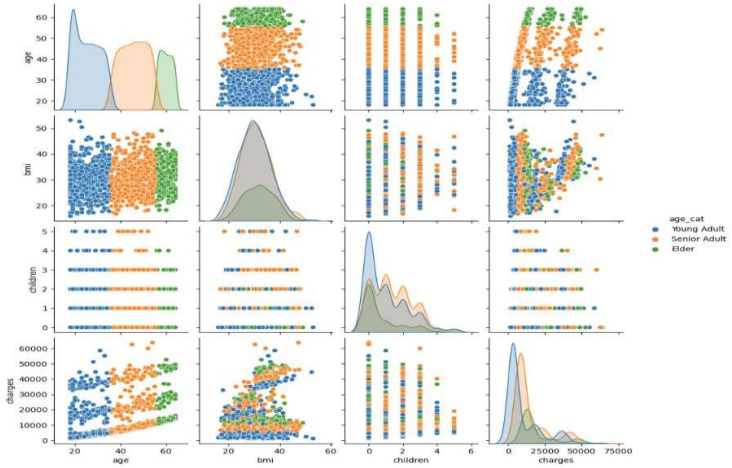
- **SHADE BY SEX**



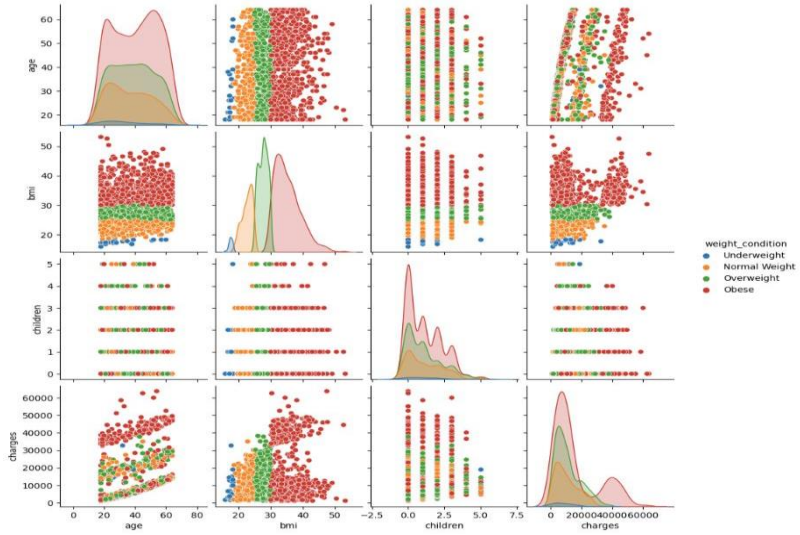
- **SHADE BY REGION**



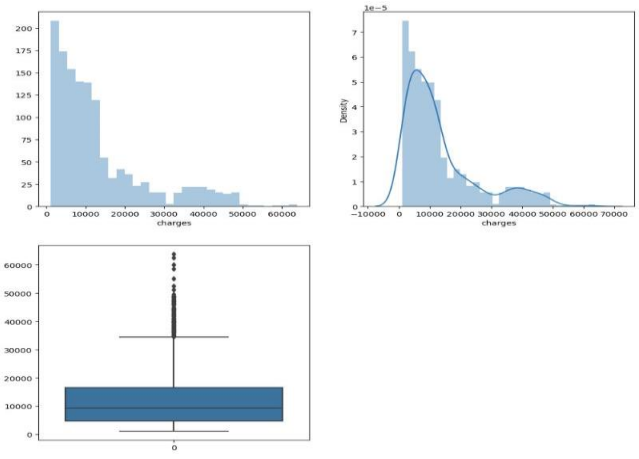
- **SHADE BY AGE**



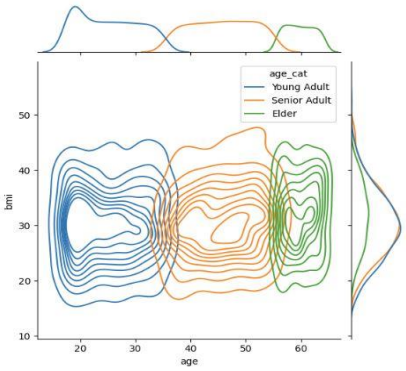
- **SHADE BY WEIGHT**

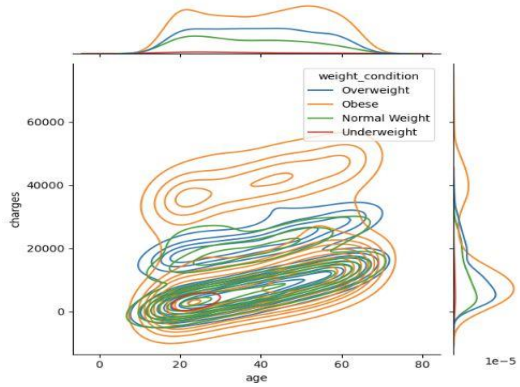


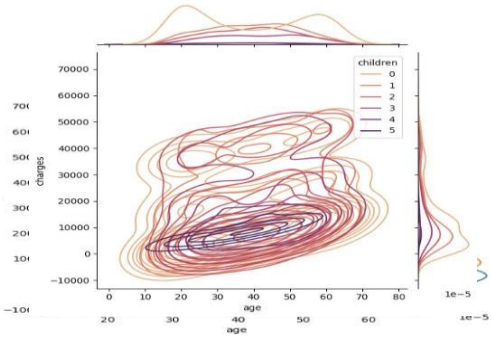
LABEL DATA



LET’S SEE SOME JOINTPLOT CHARTS



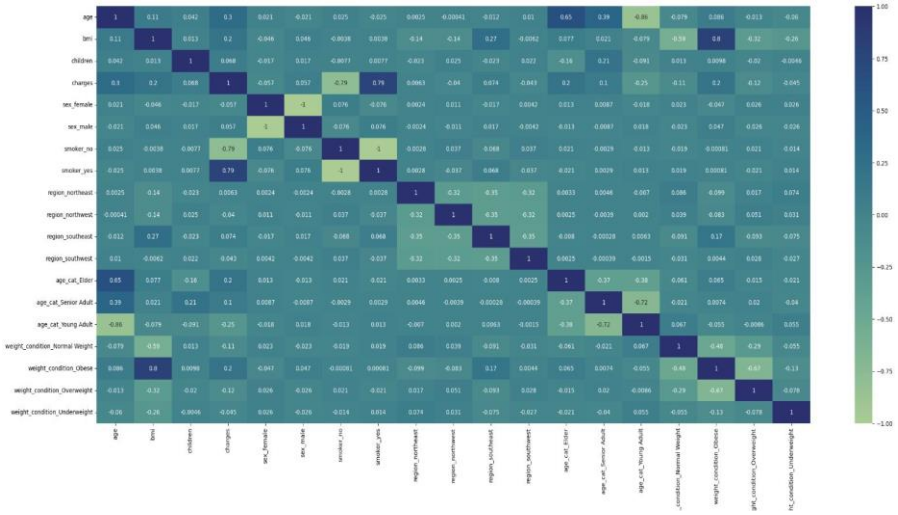




IT’S TIME FOR THE ONE HOT ENCODING

	age	bmi	children	charges	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	region_southwest	age_cat_Elder	age_cat_Senior Adult	age_cat_Young Adult	weight_condition_Normal Weight	weight_condition_Obese	weight_condition_Overweight	weight_condition_Underweight
0	19	27.900	0	16884.92400	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0
1	18	33.770	1	1725.55230	0	1	1	0	0	0	1	0	0	0	1	0	1	0	0
2	28	33.000	3	4449.46200	0	1	1	0	0	0	1	0	0	0	1	0	1	0	0
3	33	22.705	0	21984.47061	0	1	1	0	0	1	0	0	0	0	1	1	0	0	0
4	32	28.880	0	3866.85520	0	1	1	0	0	1	0	0	0	0	1	0	0	1	0

CORRELATION MATRIX



SPLITTING THE DATA

```
# splitting data into features X, and labels y
X = med_cost.drop(['charges','age','bmi'], axis=1)
y = med_cost['charges']

# splitting data into train and test data
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,
                                                    random_state=42)

# scaling values
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
X_train
```

DIFFERENT REGRESSION MODELS

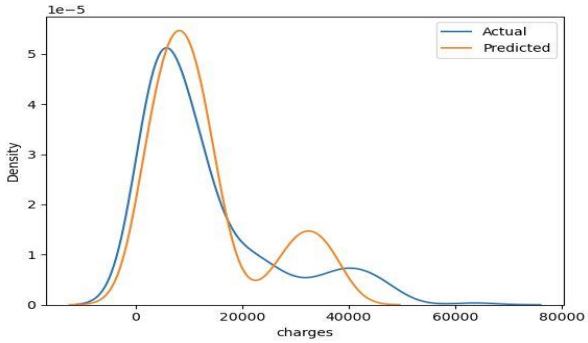
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

I decided to check what will be the results using different regression algorithms.To pick optimal hiperparmeters I used the GridSearchCV() method from scikit-learn library.

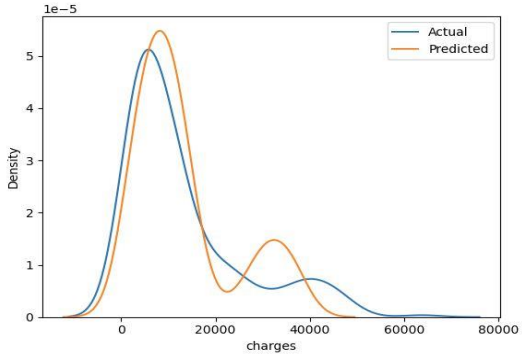
```
models = ['Linear Regression', 'Lasso Regression', 'AdaBoost Regression',
          'Ridge Regression', 'RandomForest Regression',
          'KNeighbours Regression', 'SVR']
```

LETS CHECK WHICH ALGORITHM IS THE BEST FOR THIS DATASET

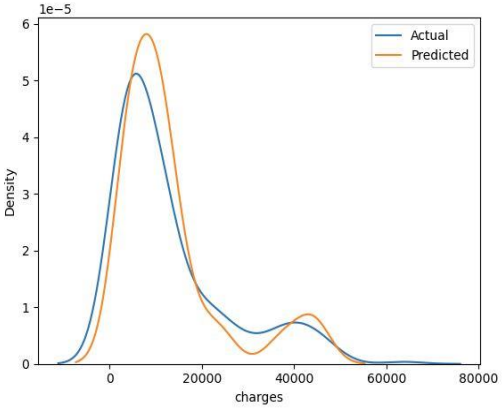
- **Liner regression**



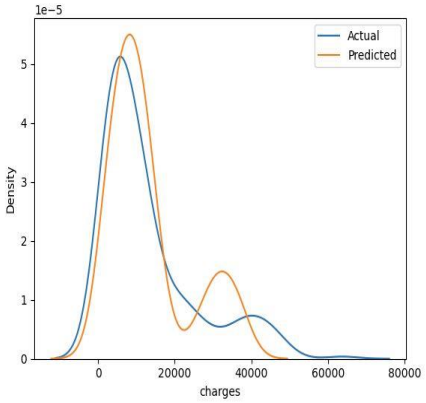
- **Lasso regression**



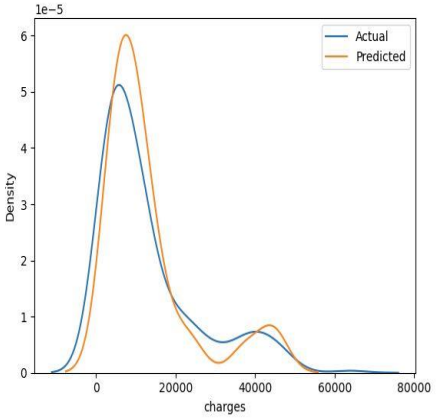
- **Adaboost regression**



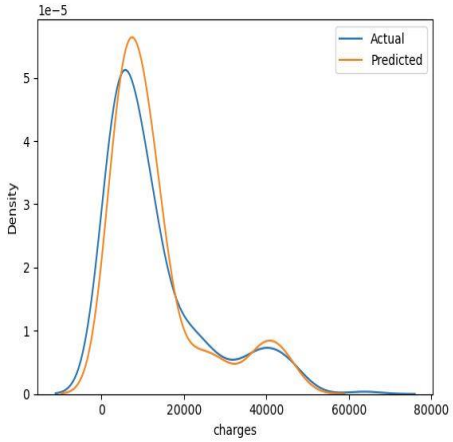
- **Ridge regression**



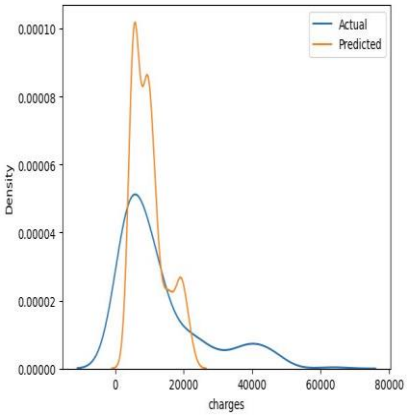
- **Random regression**



- **KNeighbours regression**



- **SVR**

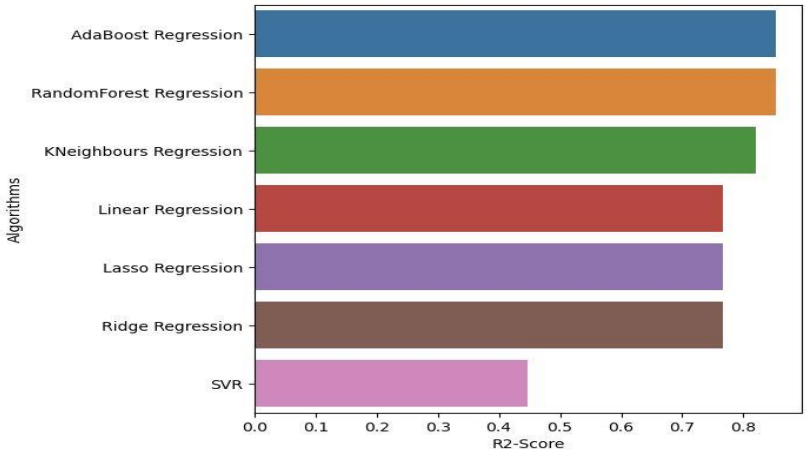


RANKING TABLE

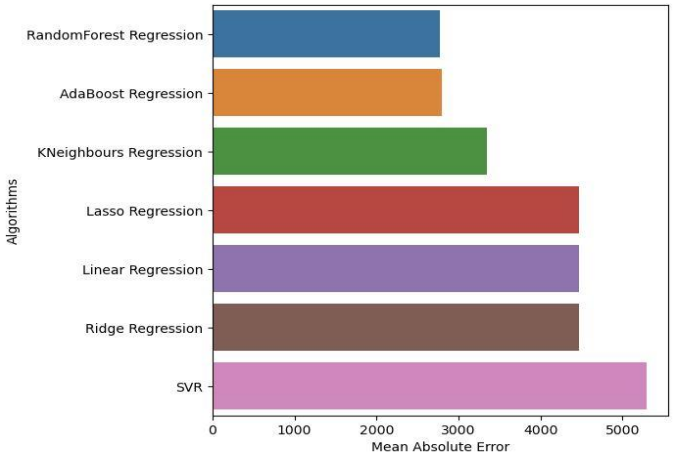
	Algorithms	R2-Score	Mean Absolute Error
2	AdaBoost Regression	0.852672	2796.002363
4	RandomForest Regression	0.852477	2776.077642
5	KNeighbours Regression	0.819795	3341.948128
0	Linear Regression	0.767019	4468.395708
1	Lasso Regression	0.766933	4467.118953
3	Ridge Regression	0.766669	4471.616012
6	SVR	0.447384	5291.739587

BARKOT CHARTS

- **Coefficient of determination -R2-score**



- **Mean absolute error – MAE**



SUMMARY

Estimating and predicting hospitalization and medical care costs can be complex and depend on various factors, such as the type of medical condition, duration of hospital stay, required treatments, and healthcare policies. Advanced statistical models and data analysis techniques are typically used for accurate predictions.

To get a summary estimate, you can consider using historical data and regression models to identify patterns and correlations between medical conditions, hospitalization, and costs. Additionally, machine learning algorithms, like decision trees or neural networks, can also be employed to improve accuracy. Keep in mind that any prediction or estimation is only as good as the data it's based on. The more comprehensive and recent the data, the more accurate your

predictions are likely to be. Always verify your findings with healthcare professionals and consult experts for a more precise understanding of hospitalization and medical care costs in specific scenarios.

DECLARATION

To estimate and predict hospitalization and medical care costs, you would typically need historical data, relevant variables, and a suitable model. These costs can be influenced by various factors like the type of treatment, duration of hospitalization, patient demographics, and medical conditions. If you have access to the necessary data, statistical techniques such as regression analysis or machine learning algorithms can be employed to build predictive models. These models can then be used to estimate future costs based on given parameters.

Keep in mind that accurate predictions depend on the quality and quantity of available data, as well as the complexity of the underlying factors influencing the costs. Additionally, as an AI language model, I don't have access to real-time data, so I can't perform specific predictions on your behalf. If you need help with a specific dataset or model, feel free to provide more details, and I'll do my best to assist you further.

CONCLUSION

Estimating and predicting hospitalization and medical care costs is a complex process that involves analyzing various factors such as the patient's medical history, severity of the condition, treatment plans, duration of hospital stay, and regional healthcare costs. Sophisticated statistical models and machine learning algorithms can be utilized to make accurate predictions.

In conclusion, accurately estimating and predicting hospitalization and medical care costs can be crucial for healthcare providers, insurers, and patients to plan and manage resources effectively. However, it's essential to continuously update and refine these models as new data becomes available to ensure their accuracy and relevance.