AI-Powered Insurance Premium Estimator

Dynamic Pricing Model with CSV Integration

Ministry of Youth and Sports
AMG, ICAIL

AMG
A Multi Group

جمهورية مصر العربية
وزارة الشباب والرياضة
Ministry of Youth and Sports

ICAIL
International Certification of Artificial Intelligence License

AI-Powered Insurance Premium Estimator

Dynamic Pricing Model with CSV Integration

In partial fulfilment of the requirements for the license of

International Certificate of Artificial Intelligence.

Dr. Soha Farahat
Associate professor of Immunology,
AI Practitioner, ICAIL Program
(2025-icail-1233)

Omar Medhat
Computer and Communication Engineering Student at Alexandria University,
AI Practitioner, ICAIL Program
(2025-icail-1301)

Yasmine Mostafa
Actuarial Science Student at Cairo University,
AI Practitioner, ICAIL Program
(2025-icail-1356)

Contents

Project Overview

Our project developed a **machine learning system** to predict insurance premiums, and was built using **historical data, customer demographics, and medical risk factors**.

The system includes:

- A **CSV analyzer** for uploading datasets and running bulk predictions.
- A **trained predictive model**.
- An **interactive dashboard** for visualizing insights.

Problem Statement

Traditional pricing is often ***manual*** , ***very complex*** and ***time consuming***.

There’s a need in the market for:

- High-accuracy
- Data-driven predictions
- Transparency on how variables like BMI and smoking affect cost
- Real-time, bulk decision support

Goal and Objectives

Goal: Develop a machine learning-based system that accurately predicts insurance premiums using customer data.

Objectives:

- Read and parse CSV files.
- Generate statistical summaries.
- Display interactive dashboards.
- Implement an AI model for premiums prediction.
- Ensure the system is scalable and easy to update with new data or features.

Data Overview

Dataset used: Health care insurance (Kaggle) – 1338 records

Features:

- **age**: age of primary beneficiary
- **sex**: insurance contractor gender, female, male
- **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height
- **children**: Number of children covered by health insurance / Number of dependents
- **smoker**: Smoking status
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges**: Individual medical costs billed by health insurance

CSV Analyzer

```python
from kagglehub import dataset_download

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
```

```python
def dataInfo(df):
    print(df.info())
```

```
<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: 73.3+ KB
```

CSV Analyzer

```
def catEncoding(df):
```

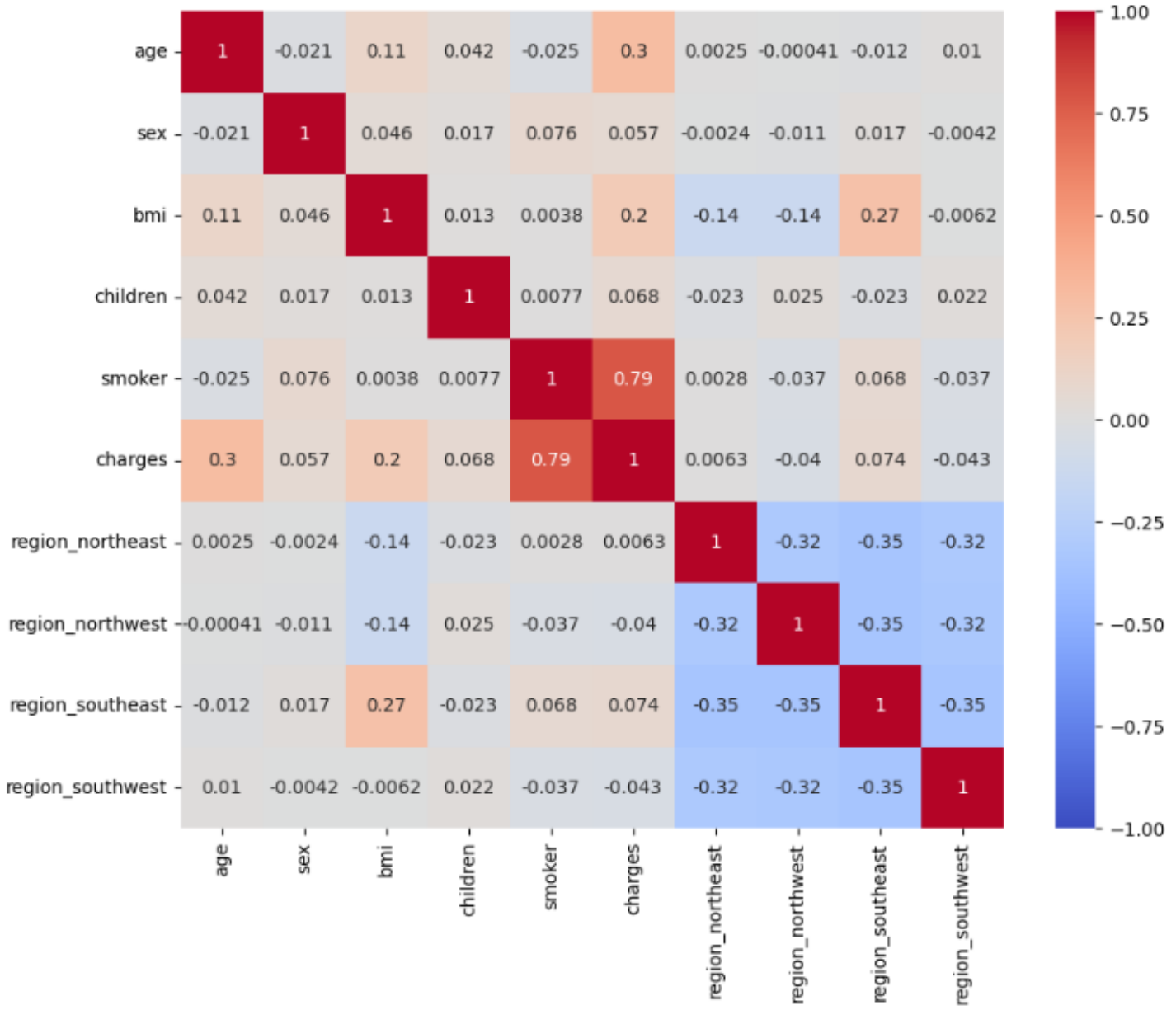
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   age                1338 non-null   int64
 1   sex                1338 non-null   int8
 2   bmi                1338 non-null   float64
 3   children           1338 non-null   int64
 4   smoker             1338 non-null   int8
 5   charges            1338 non-null   float64
 6   region_northeast   1338 non-null   int64
 7   region_northwest   1338 non-null   int64
 8   region_southeast   1338 non-null   int64
 9   region_southwest   1338 non-null   int64
dtypes: float64(2), int64(6), int8(2)
memory usage: 86.4 KB
```

```
   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
                              0     no  northwest  21984.47061
                              0     no  northwest   3866.85520
```

```
==========
es
alues
alues
==========
en   smoker      charges  region_northeast  \
 0        1  16884.92400                 0
 1        0   1725.55230                 0
 3        0   4449.46200                 0
 0        0  21984.47061                 0
 0        0   3866.85520                 0

_southeast  region_southwest
         0                1
         1                0
         1                0
         0                0
         0                0
```

CSV A

```
def dfVisual
  plt.figure
  sns.heatma
```



Model Building

1. **Problem Understanding**
2. **Data Characteristics**
3. **Model Selection According to Evaluation**

Model Building

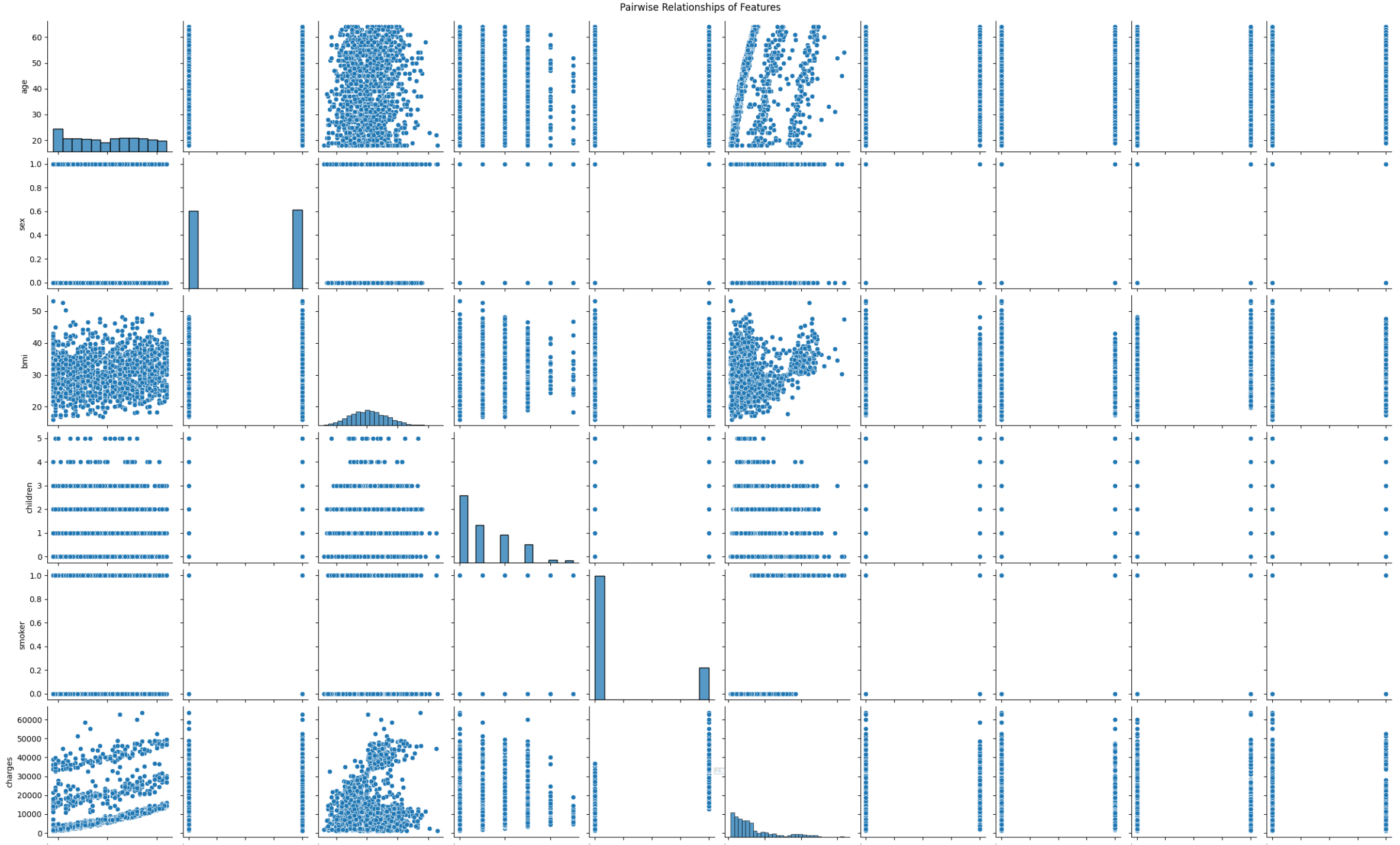
Problem Understanding

SUPERVISED LEARNING

	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

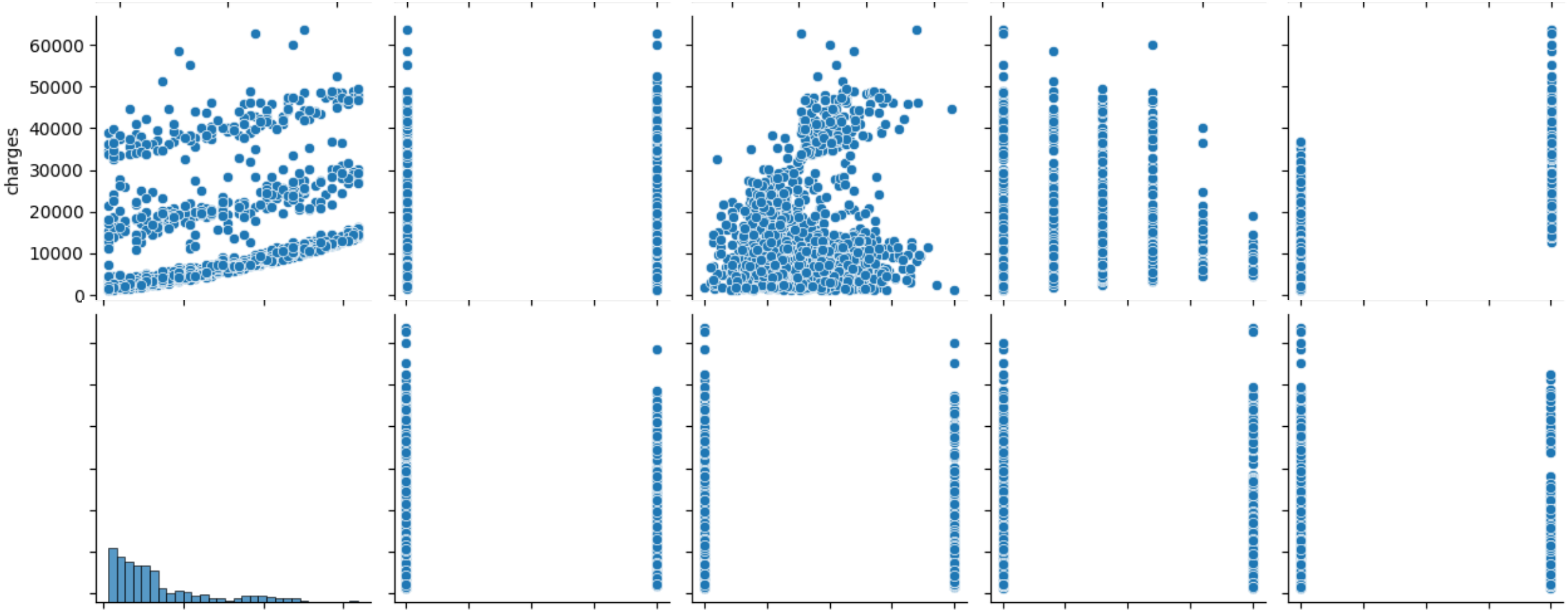
Predict insurance charges (continuous)

→ This is a regression problem


Pairwise Relationships of Features
age
sex
bmi
children
smoker
charges

Mixed features (numeric: age, bmi; categorical: sex, smoker, region)

(potential non-linear relationships, and outliers)



Model Building

Model Selection Criteria

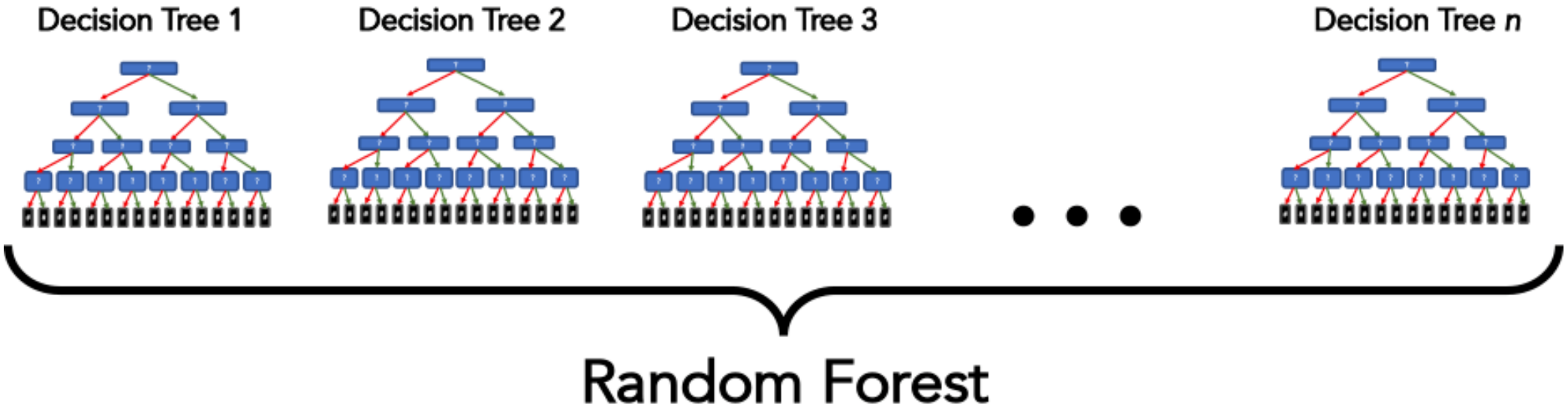
- Regression Model
- Capture non-linear relationships

Models to Evaluate

- **Linear Regression**
- **Decision Tree**

Model Building

Models Evaluated



- Addressed Decision Tree's overfitting via ensemble learning and showed better accuracy.

Random Forest Regression Model

3.1 Model Training

Random Forest Regressor

```
[ ]  x = df.drop('charges', axis=1)
     y = df['charges']
```

```
[ ]  x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.2)
```

```
[ ]  model = RandomForestRegressor()
     model.fit(x_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor()
```

3.2 Model Accuracy

```
[33] model.score(x_test, y_test)
```

```
0.8822523530318137
```

Grid Search Cross-Validation

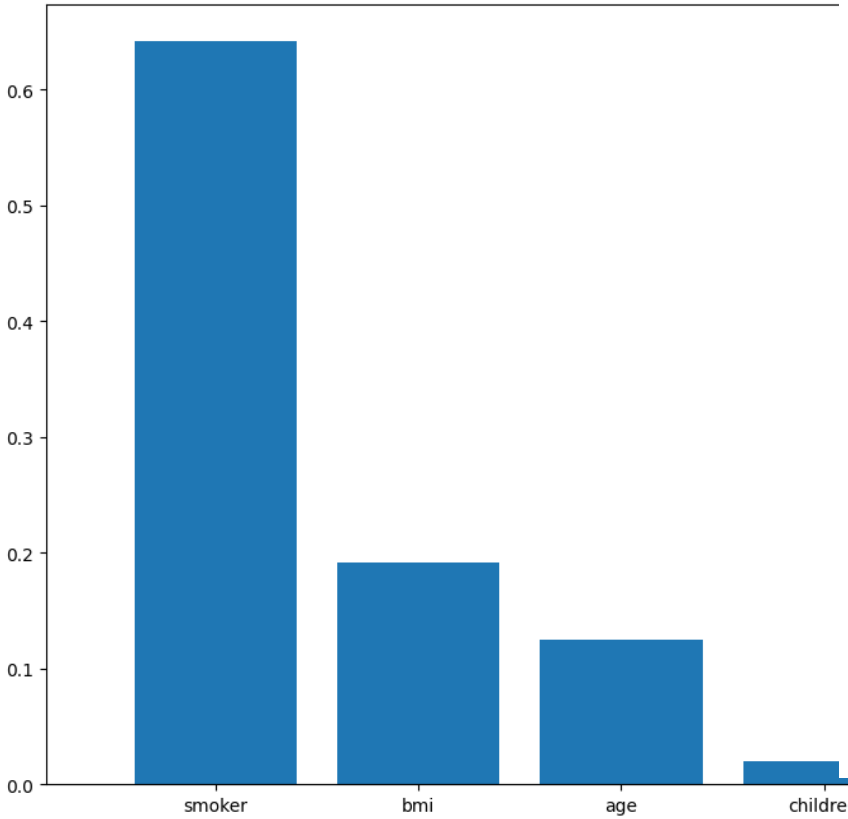
```
[28] param_
         'r
         'm
         'm
     }

     model
     grid_s
     grid_s

     print(

     Best h
```

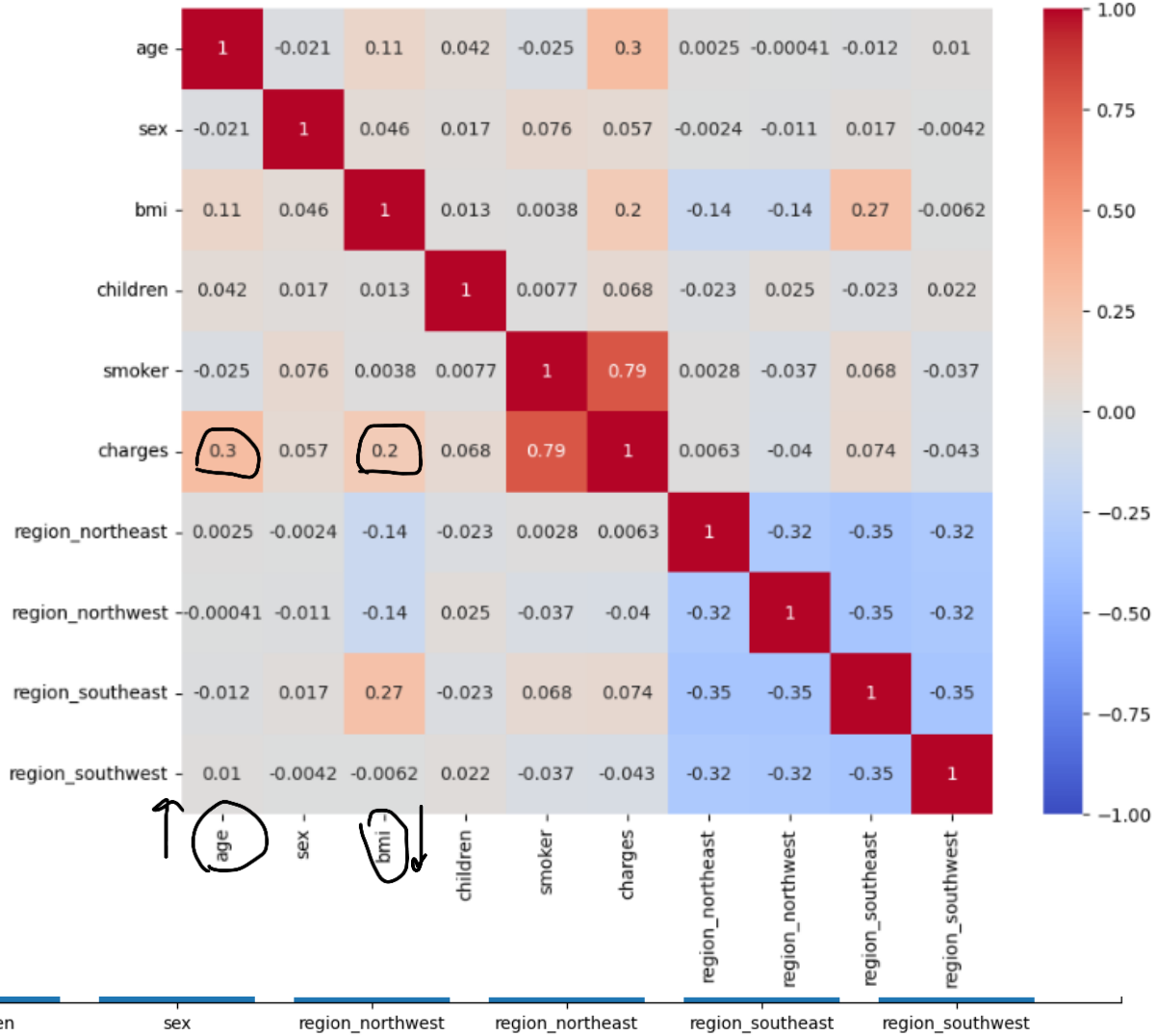
```
[29] best_params = {'max_depth': 5, 'min_samples_split': 10, 'n_estimators': 200}

     optimized_model = RandomForestRegressor(
         max_depth=best_params['max_depth'],
         min_samples_split=best_params['min_samples_split'],
         n_estimators=best_params['n_estimators'],
         random_state=42
     )

     optimized_model.fit(x_train, y_train)

     y_pred = optimized_model.predict(x_test)
```

```
[30] optimized_model.score(x_test, y_test)

     0.9011441343410856
```

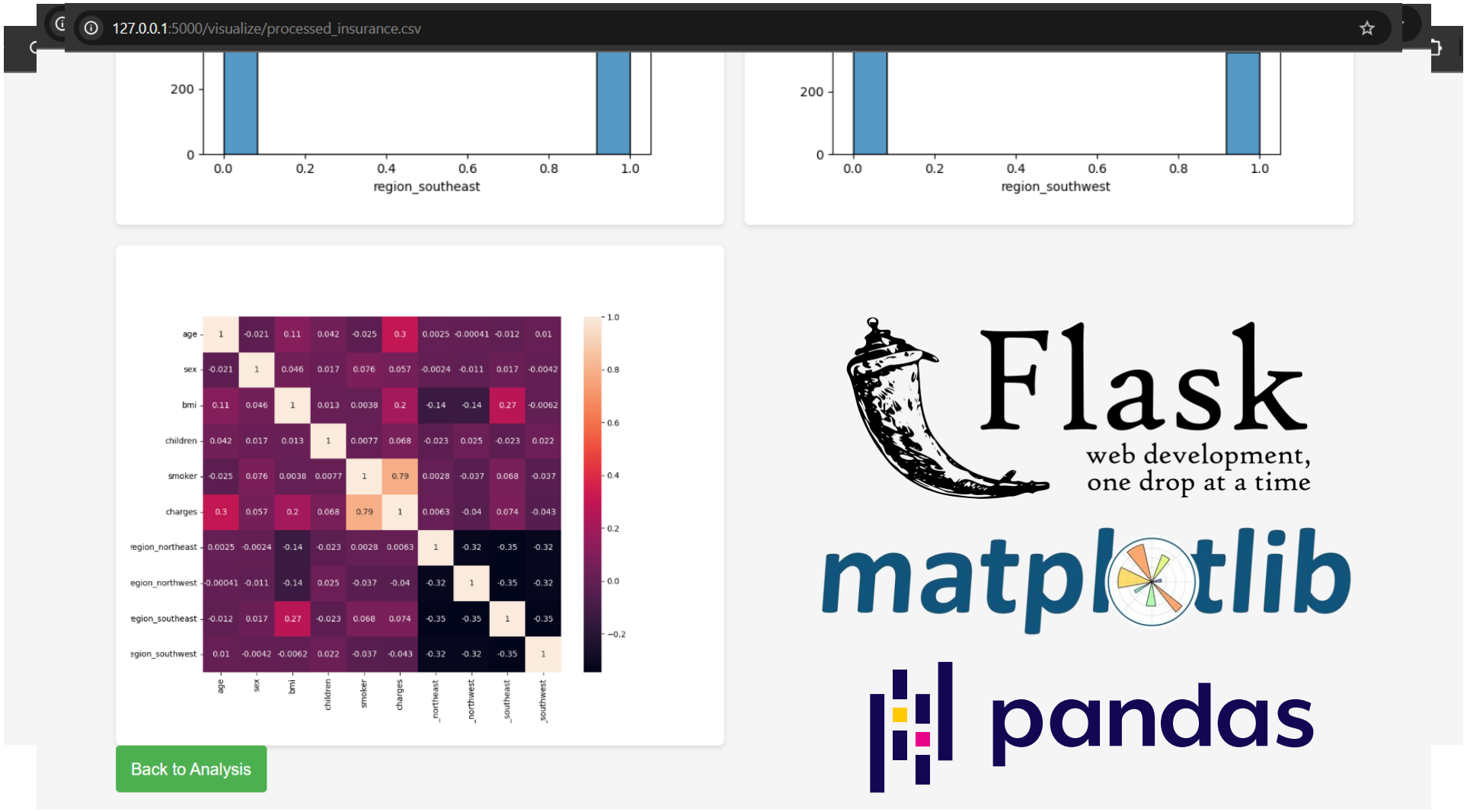
Importance Features

```
[35] feature_importances = zip(model.feature_names_i
[36]
[39]
```





Dashboard


127.0.0.1:5000/visualize/processed_insurance.csv
200
0
0.0
0.2
0.4
0.6
0.8
1.0
region_southeast
region_southwest
age
sex
bmi
children
smoker
charges
region_northeast
region_northwest
region_southeast
region_southwest
Back to Analysis
Flask
web development,
one drop at a time
matplotlib
pandas

Conclusion and Future Work

A **CSV analyzer** for uploading datasets and running bulk predictions.
A **trained predictive model using Random forest Regressor.**
A **Dashboard** for visualizing insights.

Updating the analyzer to fetch data from different sources and warehouses

GitHub Repository

Scan QR Code

GitHub

Thank You!

Any Questions?