Medical Insurance Cost Prediction

- Predicting health insurance premiums using ML and Streamlit

- Presenter: apostolosmav - https://github.com/apostolosmav

Dataset & Features

- \- Dataset Features: Age, Sex, BMI, Children, Smoker, Region, Charges

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
# Load dataset and Show first rows
df = pd.read_csv('insurance.csv')
df.head()
```

Out[2]:

	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

File display

In [3]:
```
# Check for missing values
df.isnull().sum()
```

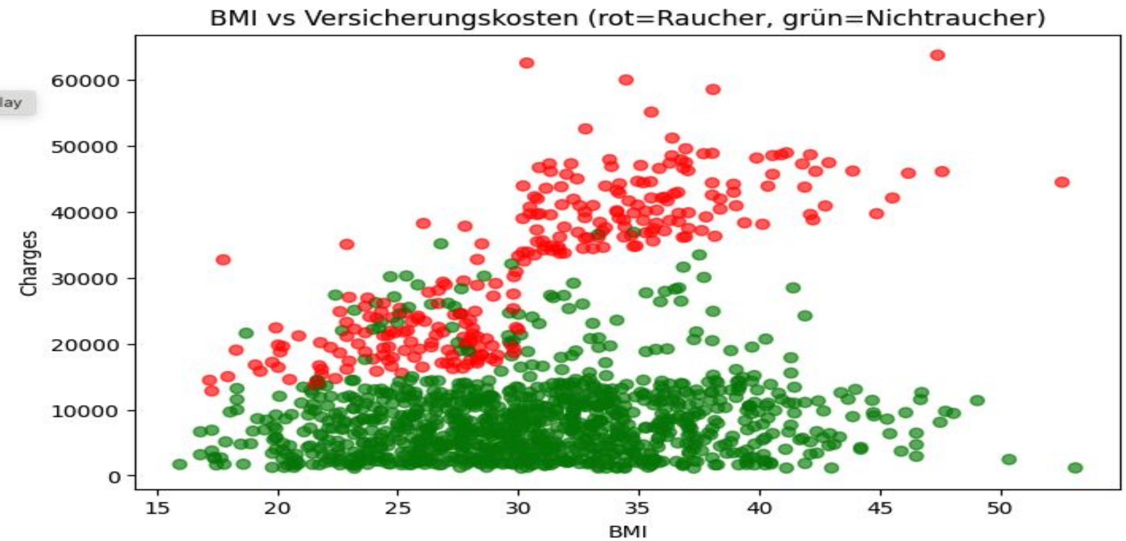
Out[3]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

In [4]:
```
# Basic statistics
df.describe()
```

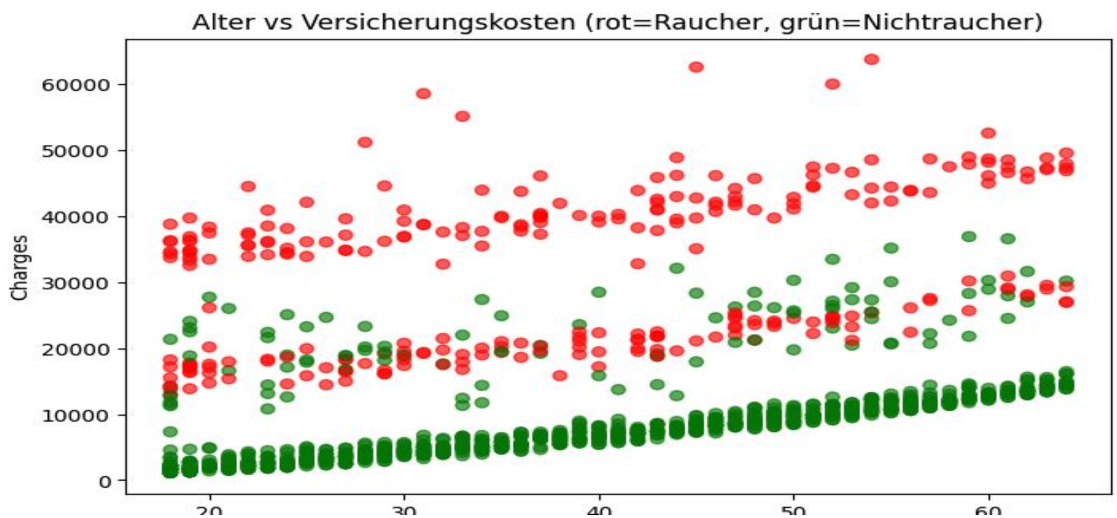
Out[4]:

	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

Exploratory Data Analysis



File display



Model Performance

- \- R^2 and MAE for all models
- \- Best model: Gradient Boosting Regressor ($R^2 \approx 0.87$)

In [ ]:

```python
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, r2_score
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor


# Prepare Data for Modeling  and Define features and target
df = pd.read_csv("insurance.csv")
X = pd.get_dummies(df.drop('charges', axis=1), drop_first=True)
y = df['charges']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train and Evaluate Multiple Models
models = {
    "Lineare Regression": LinearRegression(),
    "Decision Tree": DecisionTreeRegressor(random_state=42),
    "Random Forest": RandomForestRegressor(n_estimators=200, random_state=42),
    "Gradient Boosting": GradientBoostingRegressor(n_estimators=200, random_state=42)
}
valuation_models = {}
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    print(f"{name}: MAE={mean_absolute_error(y_test, y_pred):.2f}, R²={r2_score(y_test, y_pred):.2f}")
    valuation_models [name] = (mean_absolute_error(y_test, y_pred),r2_score(y_test,y_pred))

best_name, best_scores = max(valuation_models.items(),key= lambda x : x[1][1])

print(f"The Best model that fits these insurance Data is {best_name} with R²={valuation_models[name][1]:.2f}")

# # Train Best Model on Full Data and Save
best_model = models[best_name].fit(X, y)

import pickle as pcl

try:
    with open('Training_insurance.dat', 'wb') as file:
        pcl.dump(best_model, file)
    print("Modell erfolgreich gespeichert als 'Training_insurance.dat'")
except (IOError, FileNotFoundError):
    print("Fehler: Modell konnte nicht gespeichert werden.")
```

```
Lineare Regression: MAE=4181.19, R²=0.78
Decision Tree: MAE=3195.11, R²=0.73
Random Forest: MAE=2559.90, R²=0.86
Gradient Boosting: MAE=2492.64, R²=0.87
The Best model that fits these insurance Data is Gradient Boosting with R²=0.87
Modell erfolgreich gespeichert als 'Training_insurance.dat'
```

Technical Overview

- - Backend: Python, Scikit-learn
- - Frontend: Streamlit
- - Visualization: Matplotlib, Seaborn
- - Models: Linear Regression, Decision Tree, Random Forest, Gradient Boosting
- - Project in Notebook: https://github.com/apostolosmav/medical-insurance-analysis-prediction/blob/main/Medical_Insurance_Cost_with_Linear_Regression.ipyn

App Features

- - Inputs: Age, Sex, BMI, Children, Smoker, Region
- - Outputs: Predicted charges & interactive charts
- - Features: Dynamic sliders, real-time predictions

How It Works

- 1. User inputs data via Streamlit app (Age, Sex, BMI, etc.)
- 2. Data is preprocessed (encoding, scaling)
- 3. Trained Gradient Boosting Regressor model generates predictions
- 4. Outputs displayed with predicted charges and interactive charts

Workflow: User Input → Preprocessing → ML Model → Predictions & Charts

Exploratory Data Analysis & Insights

- Visualizations of insurance charges distribution
- - Charges by smoker status, gender, region
- - BMI vs Charges, Age vs Charges

Share

Navigation

Go to:

- EDA & Insights
- Prediction & Interaction

Exploratory Data Analysis & Insights

Dataset Preview

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.9	0	yes	southwest	16884.924
1	18	male	33.77	1	no	southeast	1725.5523
2	28	male	33	3	no	southeast	4449.462
3	33	male	22.705	0	no	northwest	21984.4706
4	32	male	28.88	0	no	northwest	3866.8552

Basic Statistics

	age	bmi	children	charges
count	1338	1338	1338	1338
mean	39.207	30.6634	1.0949	13270.4223
std	14.05	6.0982	1.2055	12110.0112
min	18	15.96	0	1121.8739
25%	27	26.2963	0	4740.2872
50%	39	30.4	1	9382.033
75%	51	34.6938	2	16639.9125
max	64	53.13	5	63770.428

Missing Values

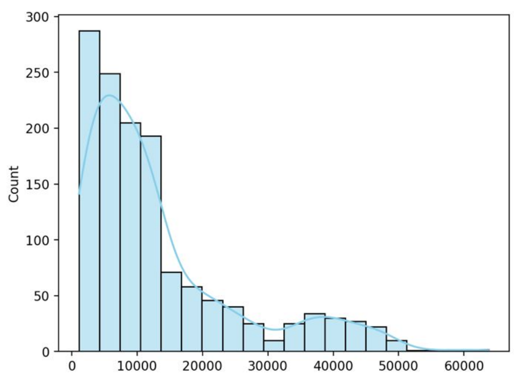
	0
age	0
sex	0
bmi	0
children	0
smoker	0
region	0
charges	0

Go to:

- EDA & Insights
- Prediction & Interaction

Manage app

Distribution of Charges



Manage app

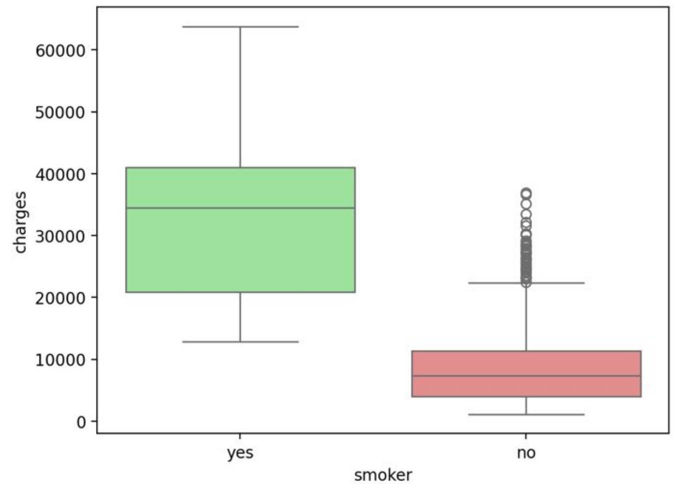
Share

Navigation

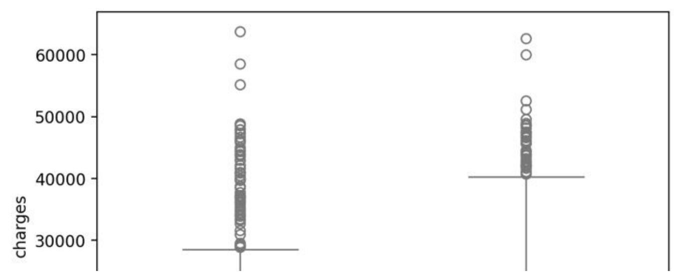
Go to:

- EDA & Insights
- Prediction & Interaction

Charts by Smoker Status



Charges by Gender



Manage app

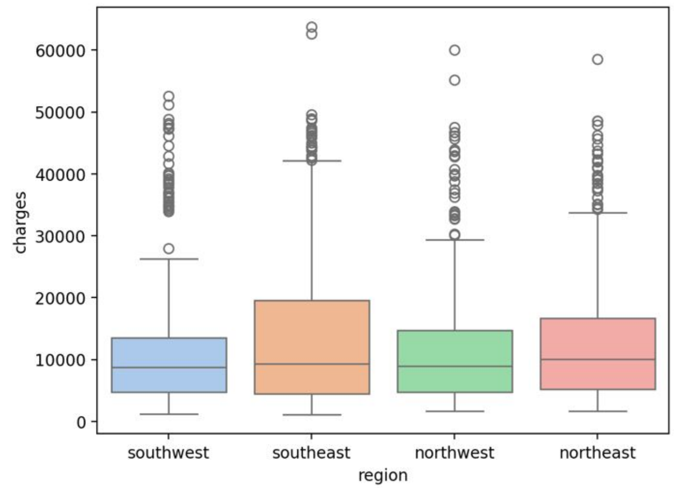
Share

Navigation

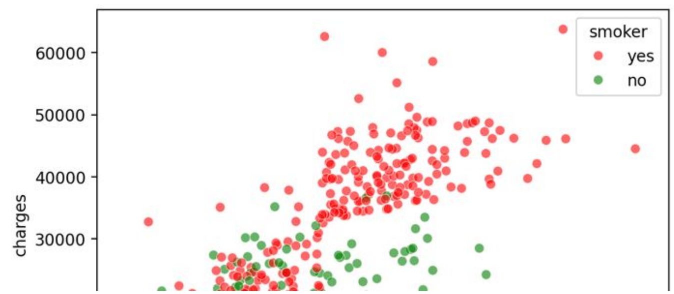
Go to:

- EDA & Insights
- Prediction & Interaction

Charges by Region



BMI vs Charges



Manage app

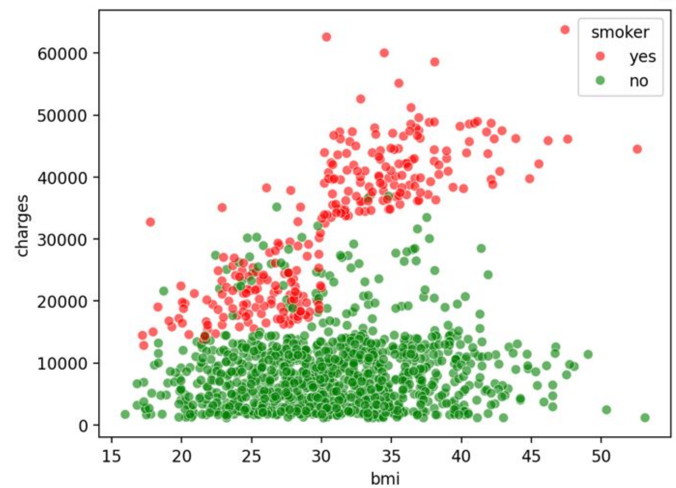
Share

Navigation

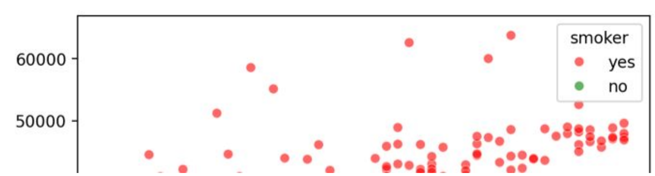
Go to:

- EDA & Insights
- Prediction & Interaction

BMI vs Charges



Age vs Charges



Manage app

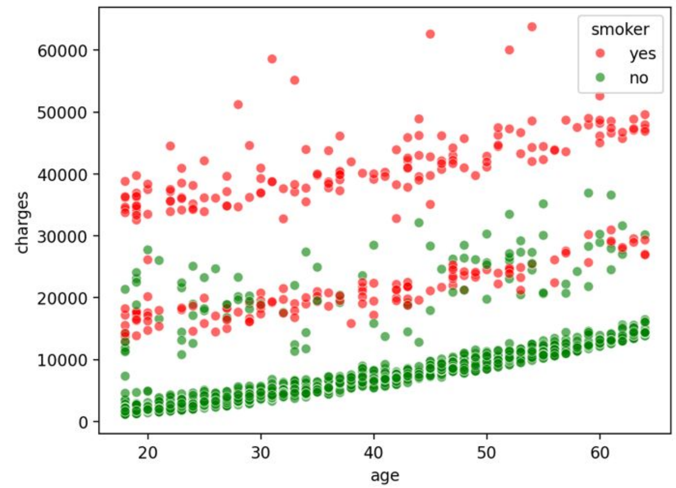
Share

Navigation

Go to:

- EDA & Insights
- Prediction & Interaction

Age vs Charges



Average Charges by Smoker and BMI Category

smoker	Underweight	Normal	Overweight	Obese
no	5485.0568	7734.6501	8226.0887	8853.2773
yes	18809.825	19942.2236	22491.1829	41692.809

Manage app

Share

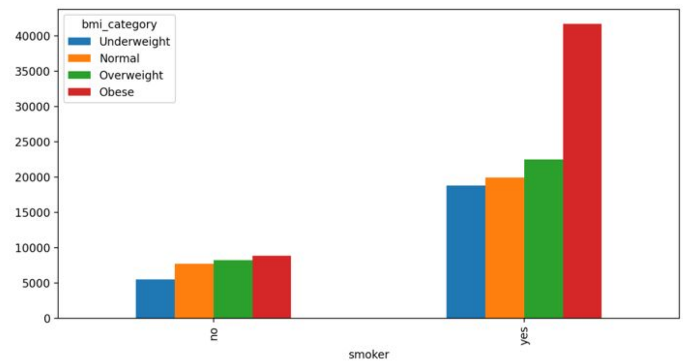
Navigation

Go to:

- EDA & Insights
- Prediction & Interaction

Average Charges by Smoker and BMI Category

smoker	Underweight	Normal	Overweight	Obese
no	5485.0568	7734.6501	8226.0887	8853.2773
yes	18809.825	19942.2236	22491.1829	41692.809



Average Charges by Smoker and Age Group

smoker	18-29	30-39	40-49	50-59	60-69	70-79	80-89	90-100
no	4418.5683	6337.3629	9183.3421	12749.3443	15232.7095	None	None	None
yes	27518.0353	30271.2464	32654.7187	37508.7529	40630.6952	None	None	None

Manage app

Share

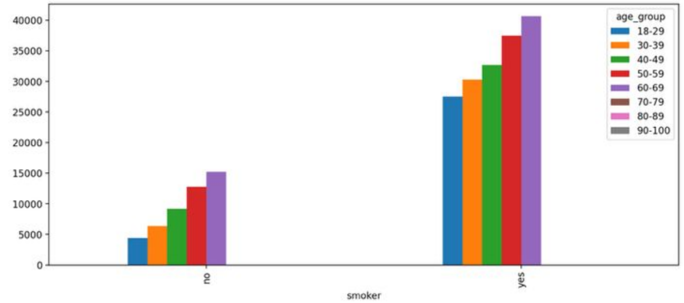
Navigation

Go to:

- EDA & Insights
- Prediction & Interaction

Average Charges by Smoker and Age Group

smoker	18-29	30-39	40-49	50-59	60-69	70-79	80-89	90-100
no	4418.5683	6337.3629	9183.3421	12749.3443	15232.7095	None	None	None
yes	27518.0353	30271.2464	32654.7187	37508.7529	40630.6952	None	None	None

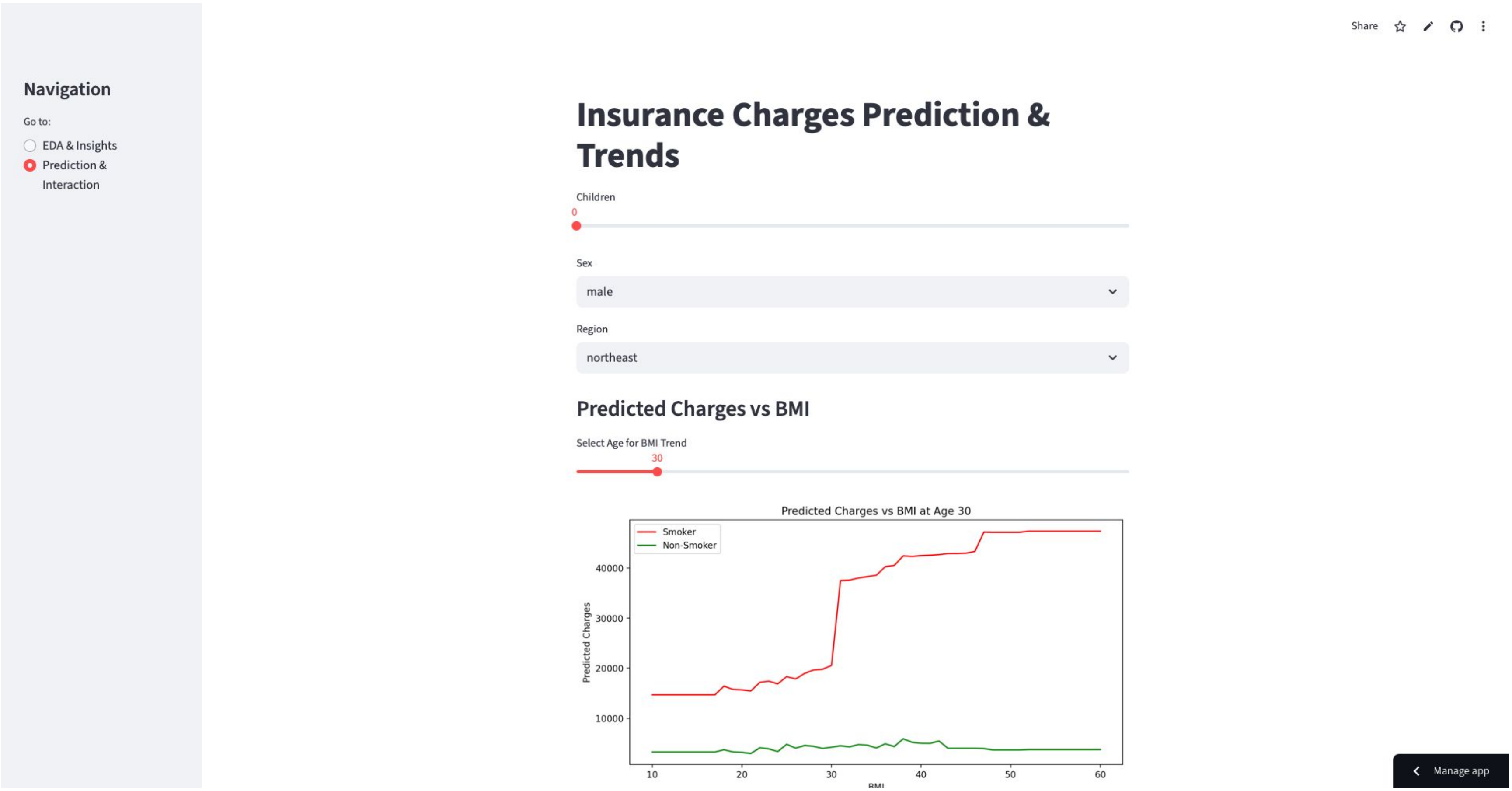


Key Insights

- **Smokers** have significantly higher insurance charges than non-smokers.
- **BMI** is strongly correlated with charges; obese smokers are the highest payers.
- **Age** increases insurance costs gradually; costs rise sharply after ~50, especially for smokers.
- **Sex** has minor effect on charges.
- **Region** has minimal impact.
- Interactions between **smoker status and BMI** or **smoker status and age** are strong drivers of charges.

Manage app

Prediction & Interaction



Share

Navigation

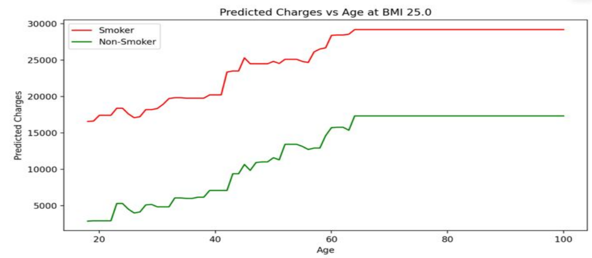
Go to:

- EDA & Insights
- Prediction & Interaction

Predicted Charges vs Age

Select BMI for Age Trend

25.00



Predict Specific Charges

Age for Specific Prediction

30

BMI for Specific Prediction

25.00

Smoker for Specific Prediction

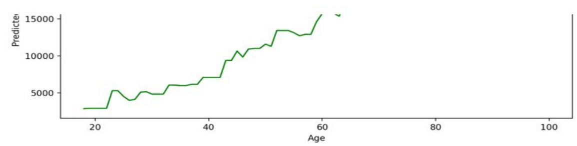
yes

Share

Navigation

Go to:

- EDA & Insights
- Prediction & Interaction



Predict Specific Charges

Age for Specific Prediction

30

BMI for Specific Prediction

25.00

Smoker for Specific Prediction

yes

Predict Specific Charges

Predicted Charges: $18347.51

Smoker: $18347.51, Non - Smoker$:4834.62

Difference (Smoker - Non-Smoker): $13512.90

Manage app

Live Demo

- - App Link:
  https://medical-insurance-analysis-prediction-rvk9rtjnciunuonjlzdbry.streamlit.app
- - Demo walkthrough: input data, view predicted charges, compare smoker vs non-smoker

Key Insights

- - Smokers pay significantly more
- - High BMI increases costs, especially for smokers
- - Costs rise with age (>50)
- - Gender & region minor effects

Future Enhancements

- - Include more features (income, pre-existing conditions)
- - Add explainable AI
- - Multi-page app with insurance plan comparison

Conclusion

- - Accurate insurance cost predictions
- - Interactive, user-friendly interface
- - Useful for individuals & insurance professionals