Medical Insurance Cost Prediction

Mini Project | Data Science Course

Prepared by: Abdullah Khatri

This project aims to predict medical insurance costs using regression models. We analyzed the dataset, performed feature engineering, applied multiple regression techniques, and tracked model performance with MLflow. A Streamlit web application was built to allow interactive visualization and cost prediction based on user input. The final deployed model helps demonstrate how data-driven approaches can enhance decision-making in healthcare insurance.

Medical insurance companies need accurate cost prediction models to estimate policyholder expenses. The goal of this project is to build a machine learning model that predicts charges based on factors such as age, sex, BMI, smoking status, children, and region. Accurate predictions can help insurers price policies fairly and allow customers to better understand their expected costs.

Data Preprocessing & Feature Engineering

- Loaded raw dataset (medical_insurance.csv).
- Handled categorical variables (sex, smoker, region → one-hot encoding).
- Created new features (BMI category, smoker-BMI interaction).
- Scaled numerical features using StandardScaler.
- Split dataset into training (80%) and testing (20%).
- Saved final cleaned dataset (cleaned_data.csv).

```
df = pd.read_csv('../data/medical_insurance.csv')
df.head()
```

✓ 0.0s

	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

```
df = pd.read_csv('../data/cleaned_data.csv')
df.head()
```

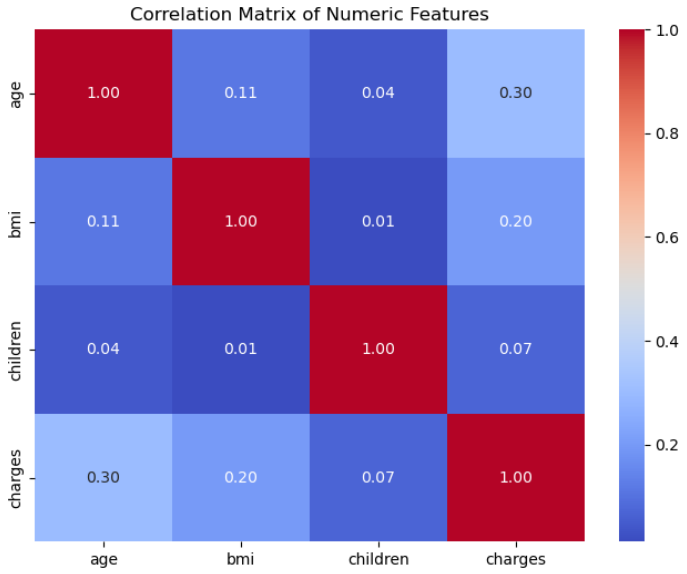
✓ 0.0s

	age	sex	bmi	children	smoker	charges	region_northwest	region_southeast	region_southwest	bmi_category_normal	bmi_category_overweight	bmi_category_underweight	smoker_bmi
0	19	0	27.900	0	1	16884.92400	0	0	1	0	1	0	27.9
1	18	1	33.770	1	0	1725.55230	0	1	0	0	0	0	0.0
2	28	1	33.000	3	0	4449.46200	0	1	0	0	0	0	0.0
3	33	1	22.705	0	0	21984.47061	1	0	0	1	0	0	0.0
4	32	1	28.880	0	0	3866.85520	1	0	0	0	1	0	0.0

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the distribution of charges, patterns, and relationships among features.

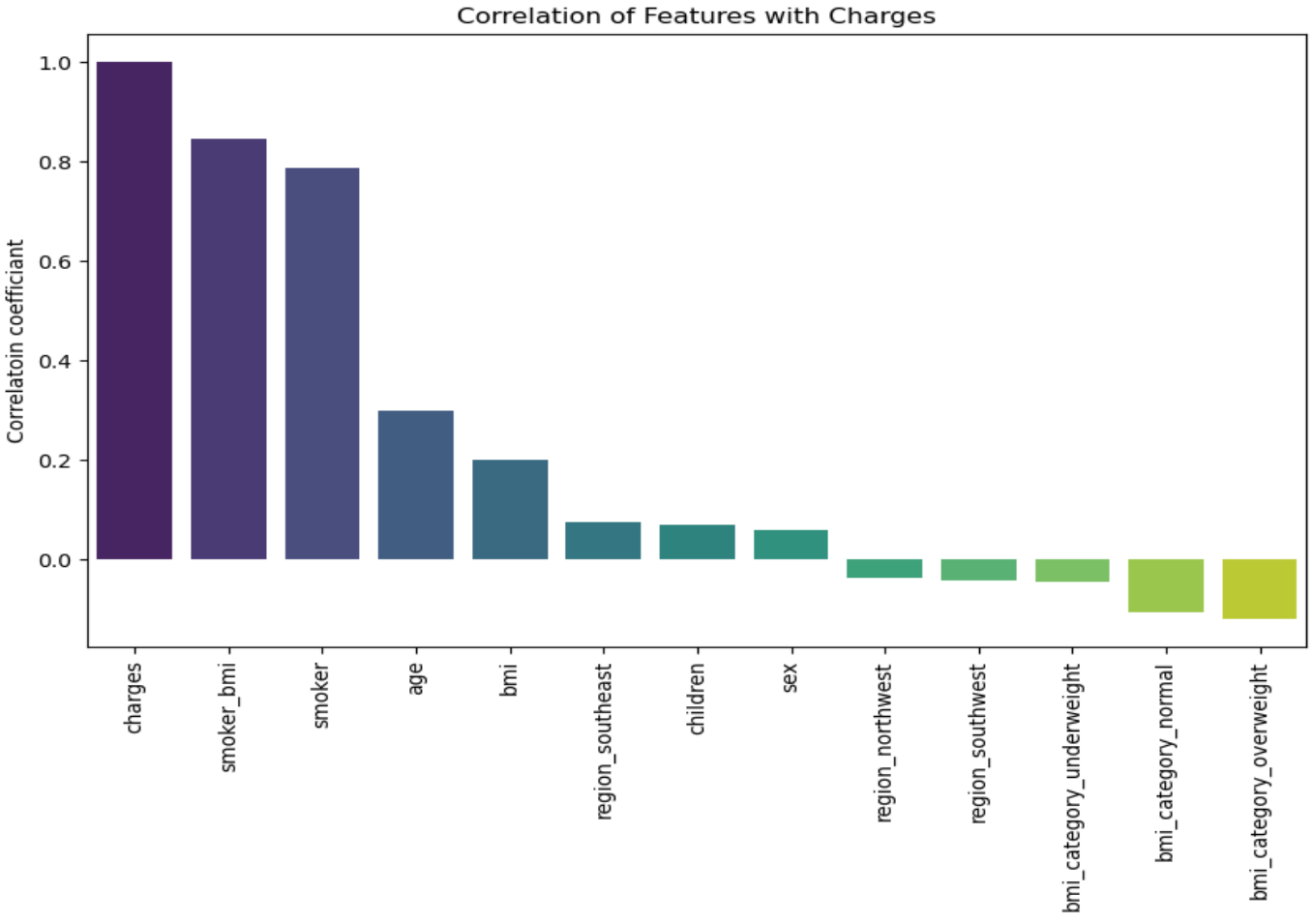
- Key insights from the EDA:
- Insurance charges increase significantly with age and BMI.
- Smokers consistently have much higher average charges than non-smokers.
- Males and females show similar charge distributions, but smoking status has a much stronger impact.
- Policyholders with more children have slightly higher average charges.
- Regional differences in charges exist but are less significant compared to smoking and BMI.



Correlation Analysis:

From the correlation heatmap and bar plot, the features that influence charges most strongly (from high to low) are:

1. **smoker_bmi**
   (strongest correlation with charges)
2. **smoker**
3. **age**
4. **bmi**
5. **children**
6. **region south_east**
7. **sex ->** *and then rest are least correlation with charges*



Model Training & MLflow Tracking

Multiple regression models were trained and evaluated to predict insurance costs.

The models included:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Polynomial Regression (degree=2)
- K-Nearest Neighbors (KNN)
- Random Forest Regressor
- XGBoost Regressor

For all models, performance was evaluated on:

i. RMSE (Root Mean Squared Error)
ii. MAE (Mean Absolute Error)
iii. R^2 Score

MLflow was integrated to track:

Model parameters (e.g., alpha for Ridge/Lasso, n_neighbors for KNN, etc.)

Performance metrics (RMSE, MAE, R^2)

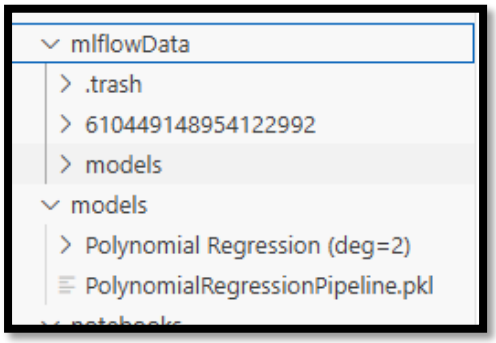
Model artifacts (saved models for future use)

The best performing model was automatically selected based on the highest R^2 score.

This model was then:

1. Registered in MLflow Model Registry under the name *MedicalInsuranceCostModel*.

2. Saved locally in the `models/` folder for use in the Streamlit app.



Streamlit Web App Development

A Streamlit web application was developed to provide an interactive interface

for insurance cost prediction and exploratory data analysis (EDA).

The app includes four main pages, accessible through a sidebar navigation:

1. **Introduction Page** -> Overview of the app's purpose and features.

2. **EDA Page** -> Interactive visualizations (Univariate, Bivariate, Multivariate, Outliers, and Correlation).

3. **Predictions Page** -> User inputs age, gender, BMI, smoking status, children, and region to get cost predictions from the trained ML model.

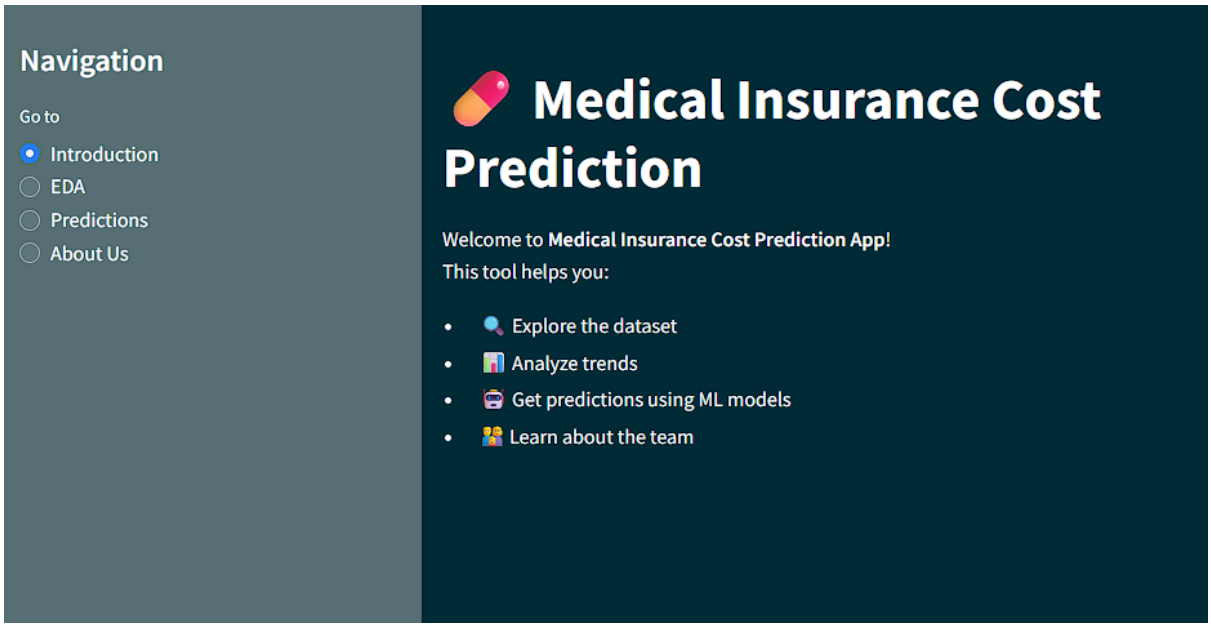
4. **About Us Page** -> Information about the project team.

The app loads the best performing model (*Polynomial Regression (deg=2)*)

from the `models/` folder and uses it to generate predictions in real time.

Additionally, a `config.toml` file was used to enhance the app's appearance

by applying a custom colour theme and layout.



Navigation

Go to

- Introduction
- EDA
- Predictions
- About Us

Exploratory Data Analysis

Select Analysis Type

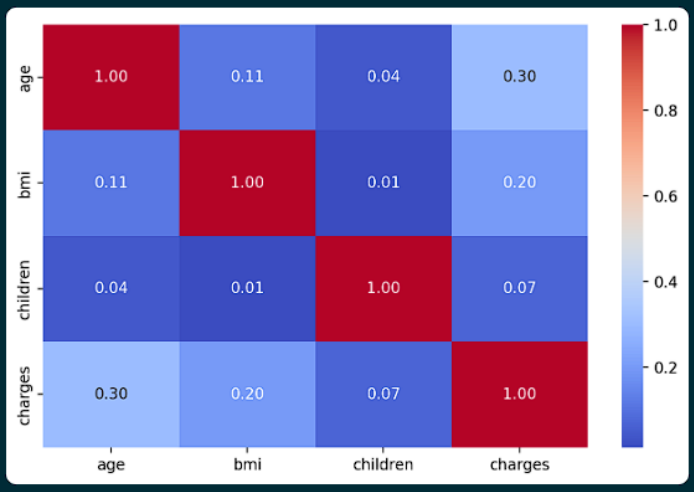
Correlation

Correlation Analysis

Select Question

Correlation Heatmap

Charges are moderately correlated with age, BMI and smoking.



Navigation

Go to

- Introduction
- EDA
- Predictions
- About Us

Exploratory Data Analysis

Select Analysis Type

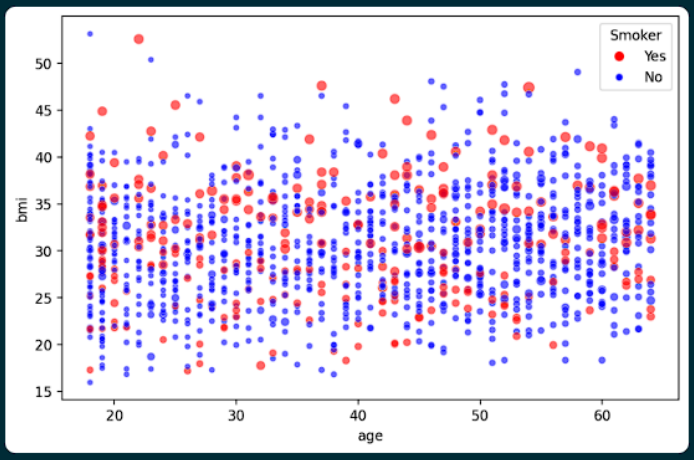
Multivariate

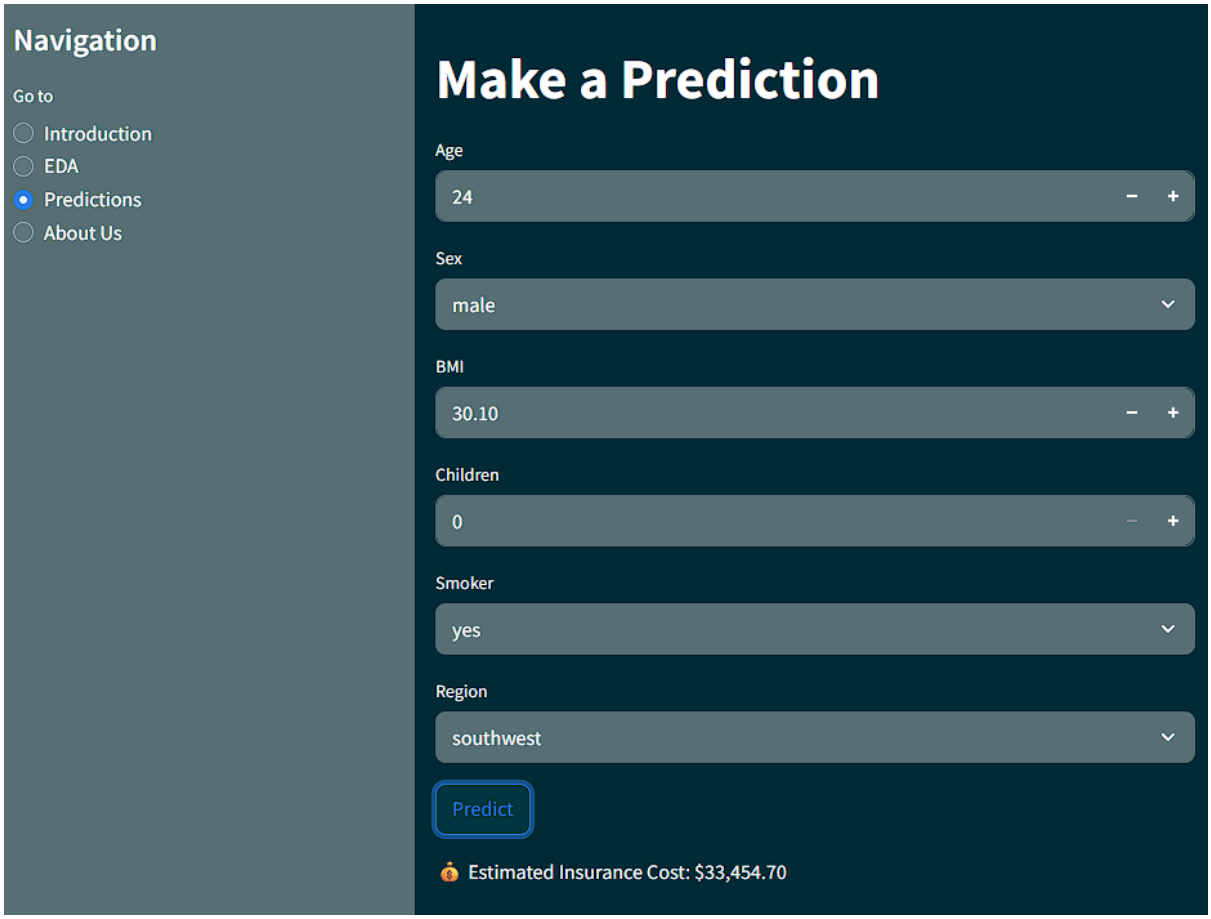
Multivariate Relationships

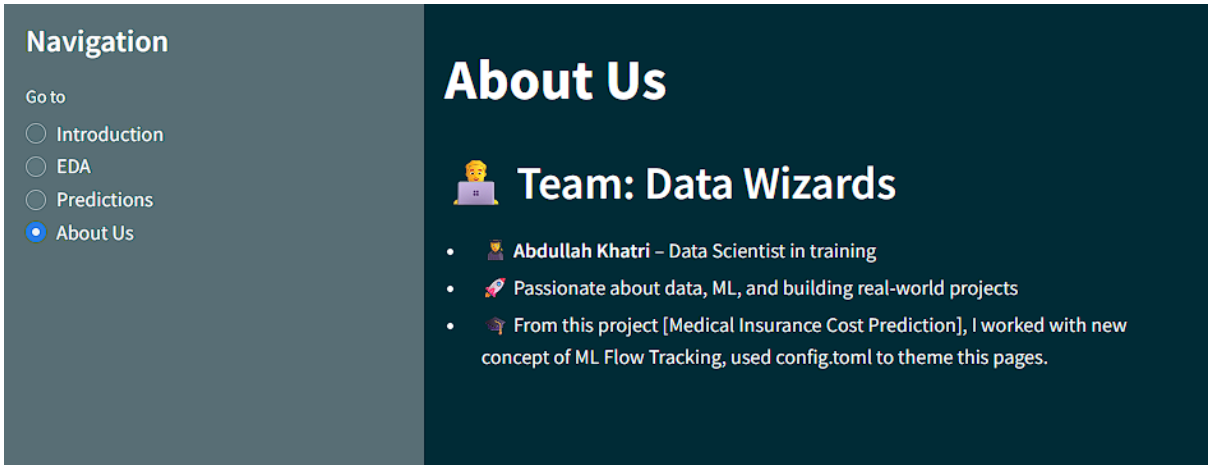
Select Question

Age, BMI and Smoking Status Impact on Charges

Smokers with high BMI and age show extreme costs.







This project successfully demonstrated the end-to-end workflow of a Data Science project:

from data preprocessing and feature engineering to model development, MLflow tracking, and deployment through a Streamlit web application.