Insurance Analysis and Streamlit Application Documentation

1. Introduction

This document provides a comprehensive explanation of the Streamlit application developed for insurance data analysis. The application, hosted on Hugging Face, leverages data science techniques to explore, visualize, and model insurance-related information. This documentation will cover data preprocessing, exploratory data analysis, model building, insights, and application deployment.

2. Understanding Insurance

Insurance is a financial arrangement that provides protection against potential losses or risks. It involves a contract between an individual (policyholder) and an insurance company, where the policyholder pays premiums in exchange for coverage against specified risks. There are various types of insurance, including health, life, auto, and property insurance.

In the context of this analysis, we focus on **health insurance**, which covers medical expenses. The premium cost is determined by several factors such as age, lifestyle choices, pre-existing conditions, and geographic location. Understanding these factors helps in predicting insurance charges effectively.

3. Data Exploration

The dataset consists of various features such as age, sex, BMI, number of children, smoking status, region, and insurance charges. The goal is to analyze the relationships between these factors and insurance costs.

Info of data

```
shape of insurance data : (1338, 7)
```

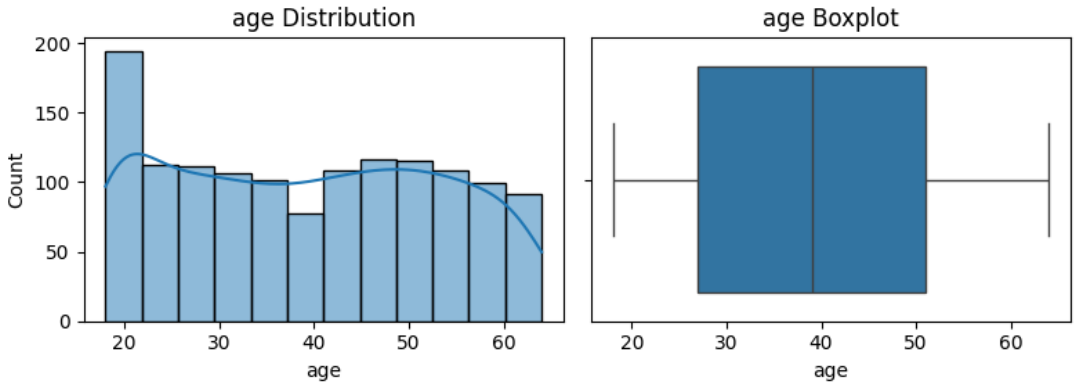
	dtype	nunique	null_count
age	int64	47	0
sex	object	2	0
bmi	float64	548	0
children	int64	6	0
smoker	object	2	0
region	object	4	0
charges	float64	1337	0

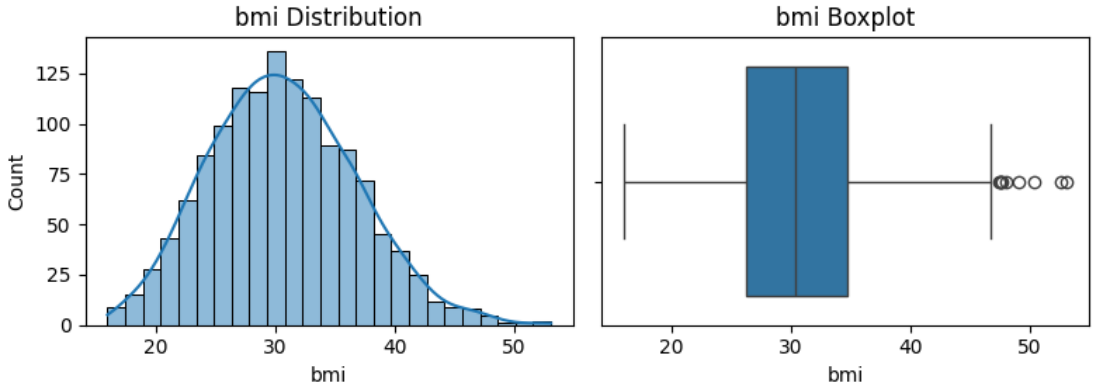
3.1 Data Preprocessing

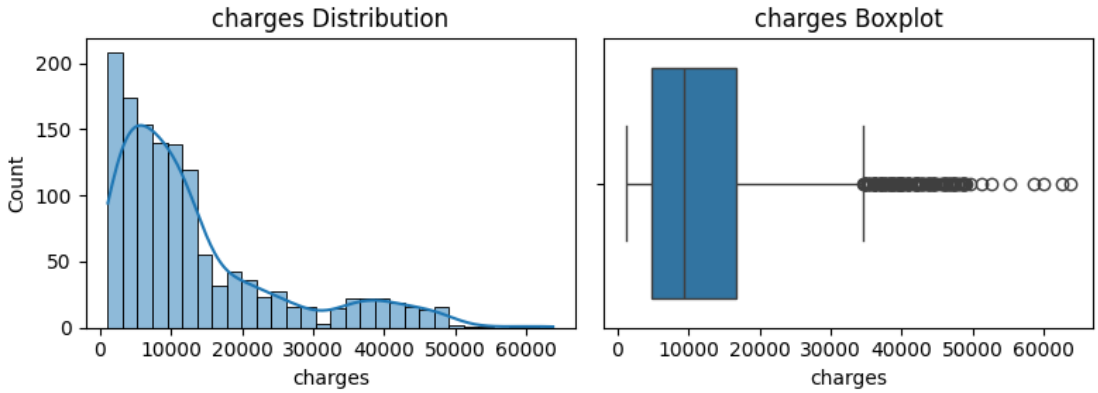
- Checking for missing values and handling them.
- Encoding categorical variables (e.g., sex, smoker, and region) using techniques like one-hot encoding or label encoding.
- Scaling numerical features where necessary.

3.2 Exploratory Data Analysis (EDA)

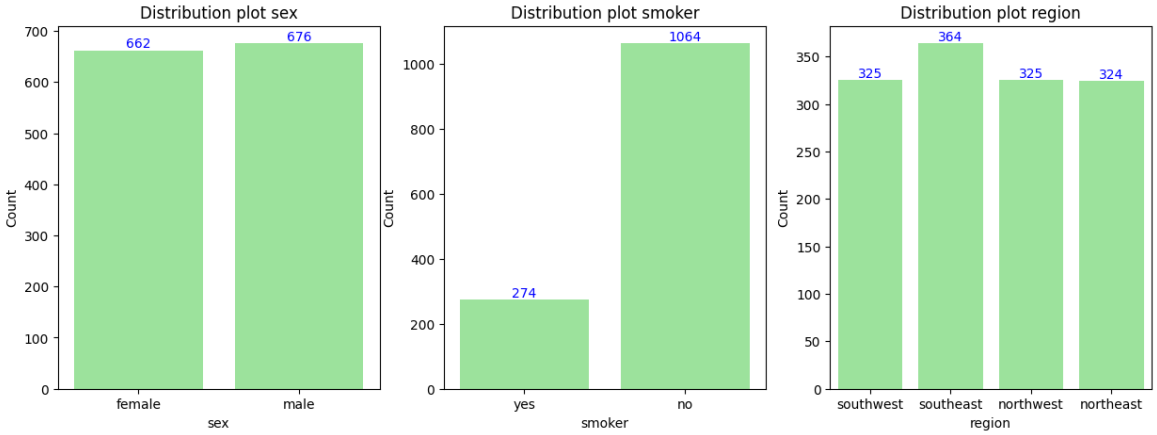
- **Distribution Analysis**: Histograms and boxplots are used to analyze the distribution of numerical variables.



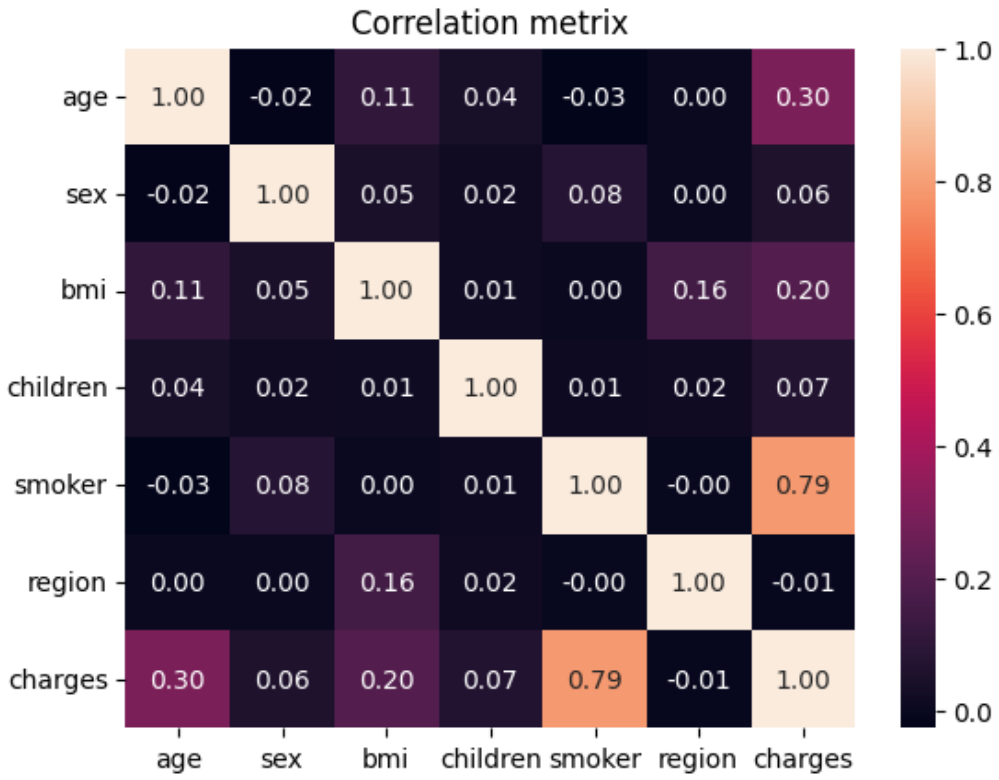




- **Barograph Analysis**: Barograph is used to analyze the distribution of catagorical variables.



- **Correlation Analysis**: A heatmap is used to examine correlations between features, identifying strong dependencies.



- **Impact of Smoking on Charges**: A boxplot helps illustrate how smoking significantly affects insurance costs.
- **Region-based Analysis**: Differences in insurance charges across different regions are examined.

4. Model Building

A machine learning model is developed to predict insurance charges based on customer attributes.

4.1 Model Selection

- **Linear Regression**: Used as a baseline model to determine relationships between independent variables and insurance charges.
- **Random Forest Regressor**: Implemented to capture non-linear relationships and improve prediction accuracy.
- **Decision Tree Regressor**: Implemented to capture non-linear relationships and improve prediction accuracy.

4.2 Model Evaluation

- **Metrics Used**:
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R-squared (R^2) score
  - Adjusted R-squared (R^2) score

Final results of models are:

	r2	rmse	mae	adj_r2
lr	5799.587091	4186.508898	0.708617	0.701918
dt	6713.129458	3043.187433	0.754057	0.748403
Rf	4635.528583	2534.285031	0.853409	0.850039

5. Streamlit Application

The interactive web application is built using Streamlit, allowing users to input parameters and obtain insurance cost predictions.

5.1 Features of the Application

- User input for age, BMI, number of children, smoking status, and region.
- Real-time prediction of insurance costs based on the trained model.
- Interactive data visualizations for better understanding.

5.2 Deployment on Hugging Face

- The application is hosted on Hugging Face Spaces, making it publicly accessible.

  Reference link

- Dependencies are managed using a requirements.txt file.

6. Insights and Findings

- **Smoking has the highest impact** on insurance costs, significantly increasing the charges.
- **Higher BMI is correlated with increased insurance charges**, emphasizing health-related risk factors.
- **Non-smokers and individuals from certain regions tend to have lower charges**, indicating potential regional influences on insurance policies.
- **Age plays a significant role** in determining insurance premiums, as older individuals generally have higher healthcare costs.

7. Conclusion

The developed Streamlit application provides an accessible and interactive way to predict insurance charges based on various customer attributes. The insights derived from the analysis can help insurance companies design better pricing strategies and assist customers in understanding their expected costs.

8. Reference

Data card:

- ✓ `https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance/data`

App:

- ✓ `https://huggingface.co/spaces/kanneboinakumar/Insurance-Analysis`

This document serves as a guide for understanding the methodology behind the insurance analysis and the implementation of the Streamlit application.