📝 Project Title: Predicting Medical Insurance Costs Using Machine Learning

---

📌 Objective

To identify key factors driving individual medical insurance costs and build a predictive model to estimate charges based on demographic and lifestyle data.

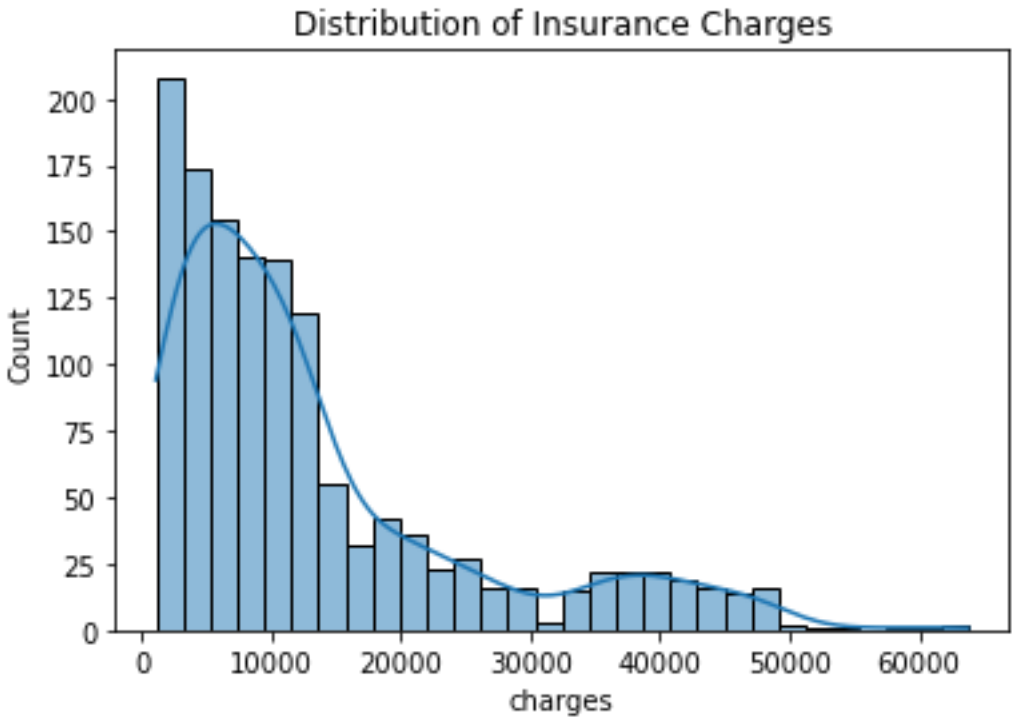
📊 Dataset

- **Source**: Kaggle – Medical Cost Personal Dataset
- **Records**: 1,338
- **Features**:
    - Age, Sex, BMI, Children, Smoker, Region, Charges

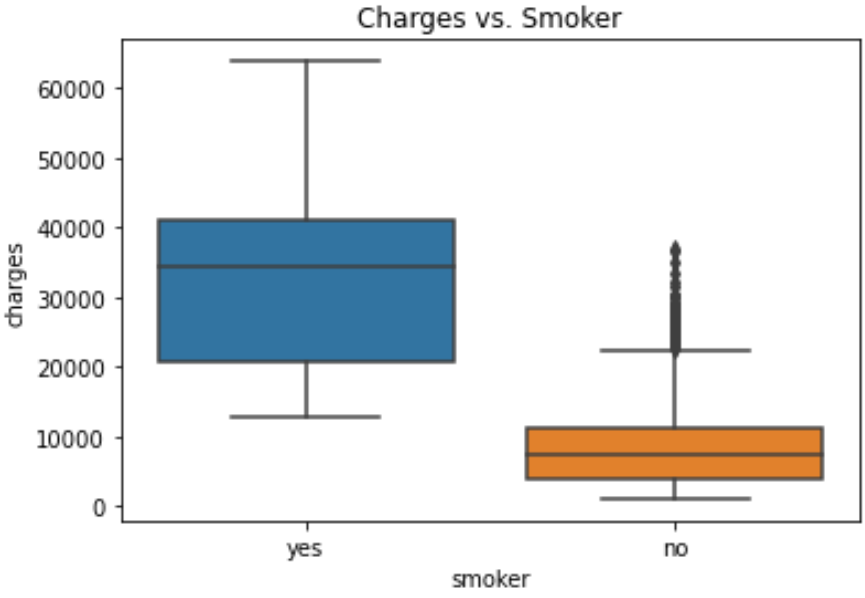
🔍 Exploratory Findings

- **Smokers** pay **~3x more** on average than non-smokers
- Charges **increase with age** and **BMI**
- Minimal cost difference across regions or gender

🔹 1. Histogram: Distribution of Charges



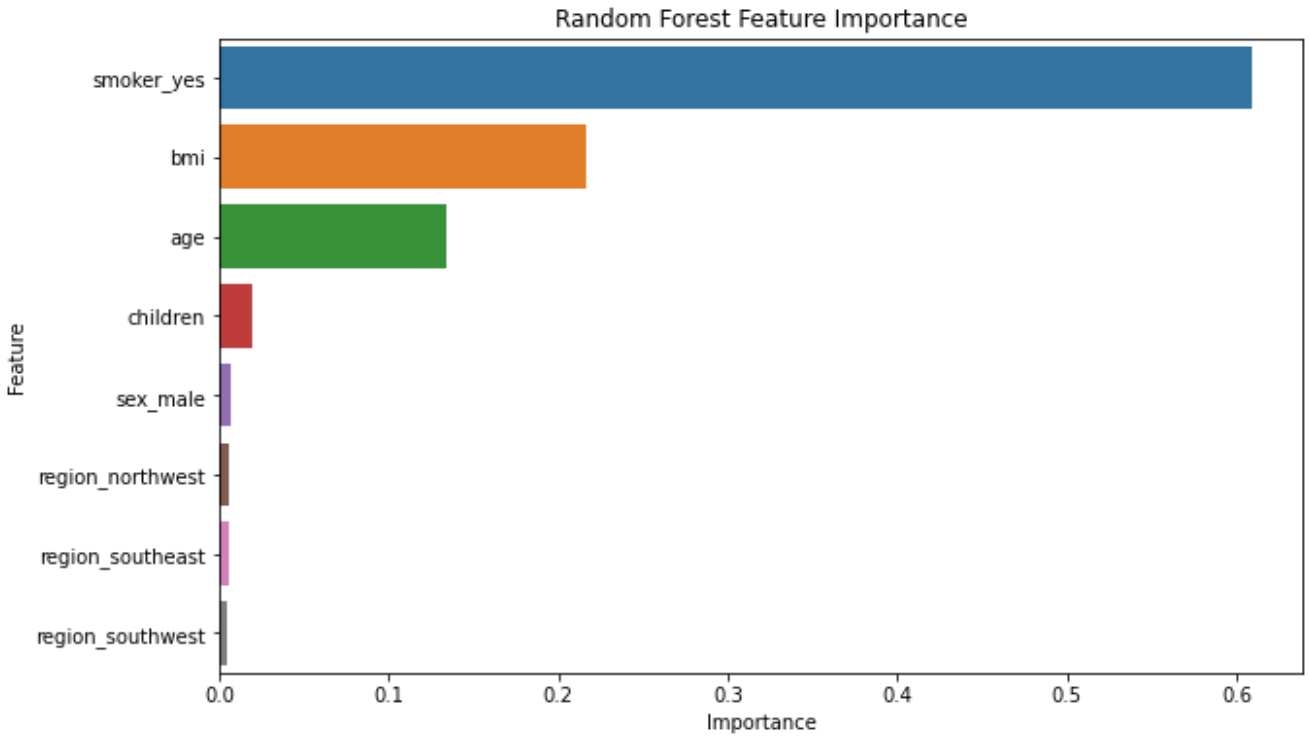
2. Boxplot: Charges vs Smoker



🤖 Modeling

Linear regression MAE is 4100, RMSE is 6100 and R^2 score is 0.79.
Random Forest MAE is 2700 , RMSE is 4200 and R^2 score is 0.88.

🔹 3. Feature Importance Chart: Random Forest



✅ Conclusion

- **Smoking is the most influential factor in insurance pricing**
- **Machine learning models can accurately predict costs from basic patient information**
- **This analysis supports pricing transparency and personalized insurance recommendations**
-

🛠️ Tools Used

- **Python (Pandas, Seaborn, Scikit-learn)**
- **Jupyter Notebook**
- **GitHub**

Project by: Saubia Aimen

🔗 Sources

- **Medical Cost Personal Dataset on Kaggle**
- **Python libraries: Pandas, Seaborn, Scikit-learn, Jupyter Notebook**