Insurance Charges Prediction Report: Regression Models Comparison

1. Objective

The goal of this analysis is to build regression models to predict individual medical charges based on demographic and health-related attributes such as age, BMI, smoking status, and region. The models used include:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor

Model performance is compared using Mean Squared Error (MSE), R^2 score, prediction vs actual visualizations, and feature importances.

2. Data Preprocessing

Initial Steps:
- Data loaded from `insurance.csv`
- No missing values were found for most features
- `autoclean()` from `datacleaner` was used for automatic preprocessing

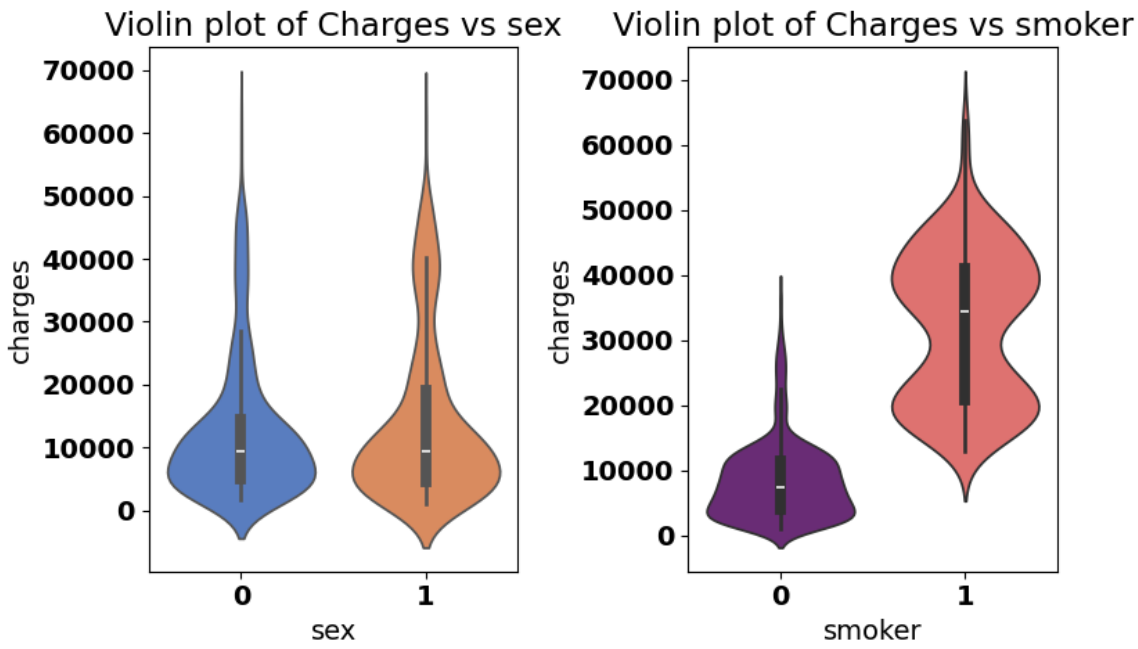
Reason for not applying manual encoding:
Categorical feature cardinality was low, and `autoclean` automatically handled this efficiently.

Feature Engineering:
- A new feature `is_risky` was added: `is_risky = smoker * bmi`
- Strong correlation observed between `is_risky` and `charges`

Scaling:
Numerical features such as age, BMI, and charges were standardized using `StandardScaler`.

3. Exploratory Data Analysis (EDA)

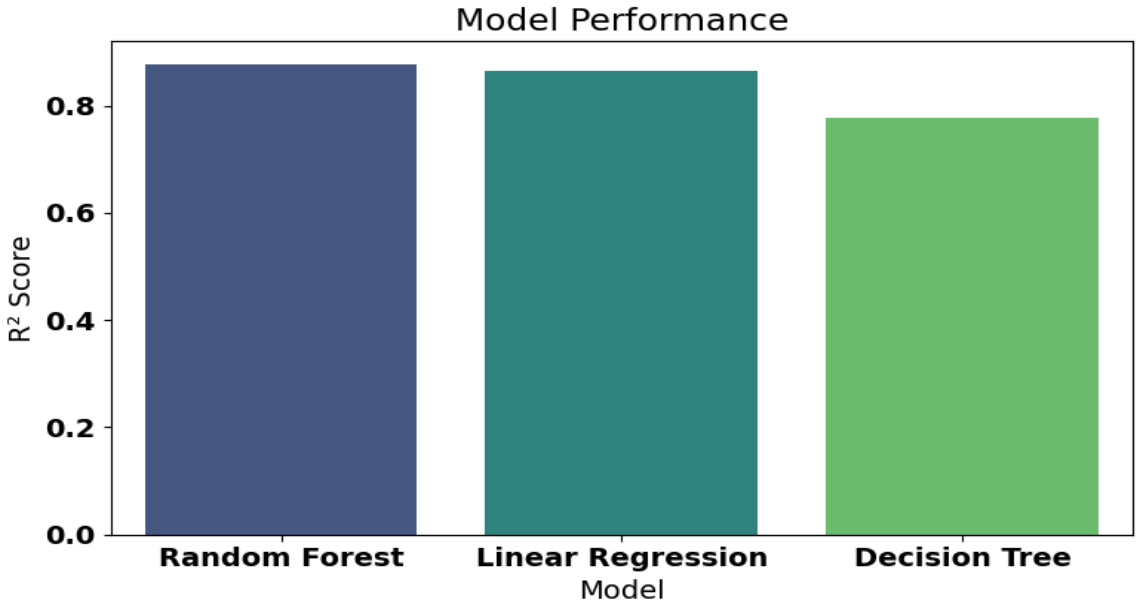
- Correlation Matrix: Showed high positive correlation between `smoker` and `charges`, and moderate correlation with `age`, `bmi`, and `is_risky`.
- Violin Plots showed:



- Smokers generally incur higher charges.
- Slightly higher charges for males.

4. Model Building & Evaluation

Models Used:
- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor



Evaluation Metrics:
- Mean Squared Error (MSE)

- R^2 Score

Results Summary:

Models Summary

Model	MSE	R^2
Random Forest	0.144493	0.873468
Linear Regression	0.246839	0.783844
Decision Tree	0.333916	0.707591

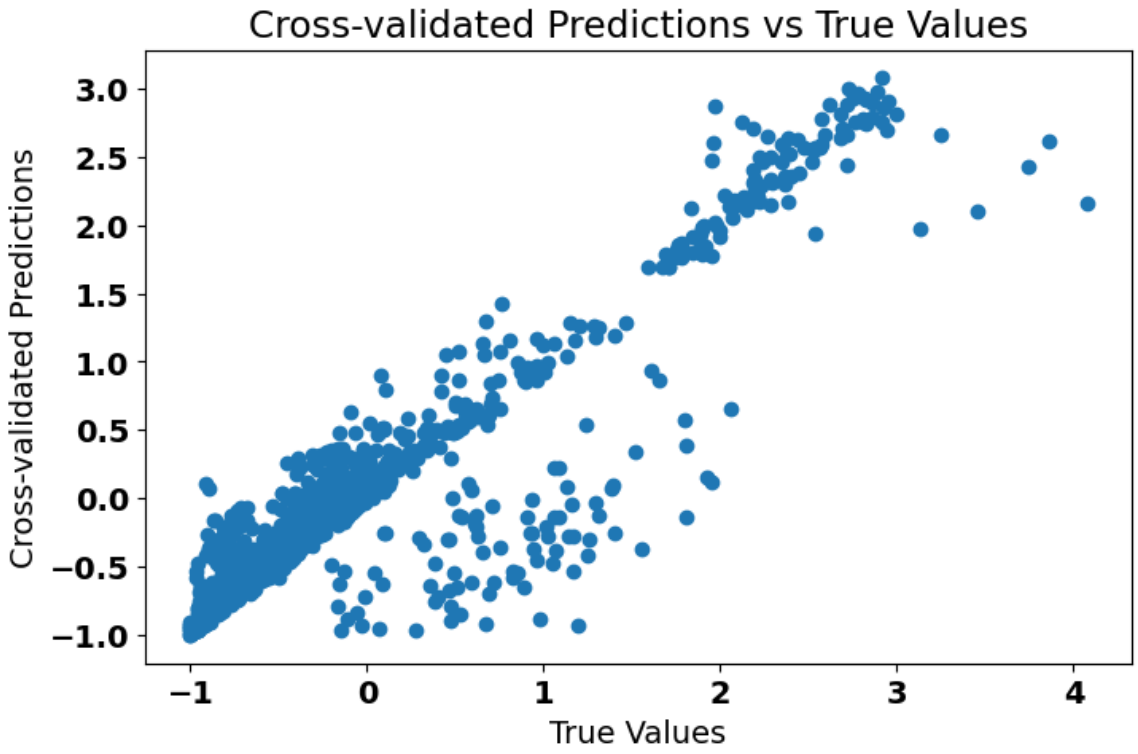
Best Model: Random Forest Regressor

5. Visualizations

- Model Performance (R^2 Score Comparison): Barplot comparing the three models
- Prediction vs Actual Charges Plots:
  - Clear alignment in Random Forest predictions
  - Linear Regression predictions slightly underperform
- Residual Plot: For further evaluation of prediction errors

6. Cross-Validation

- 5-Fold Cross-Validation was applied
- Mean R^2 Score: ~0.831
- Validated the robustness of the best model (Random Forest)

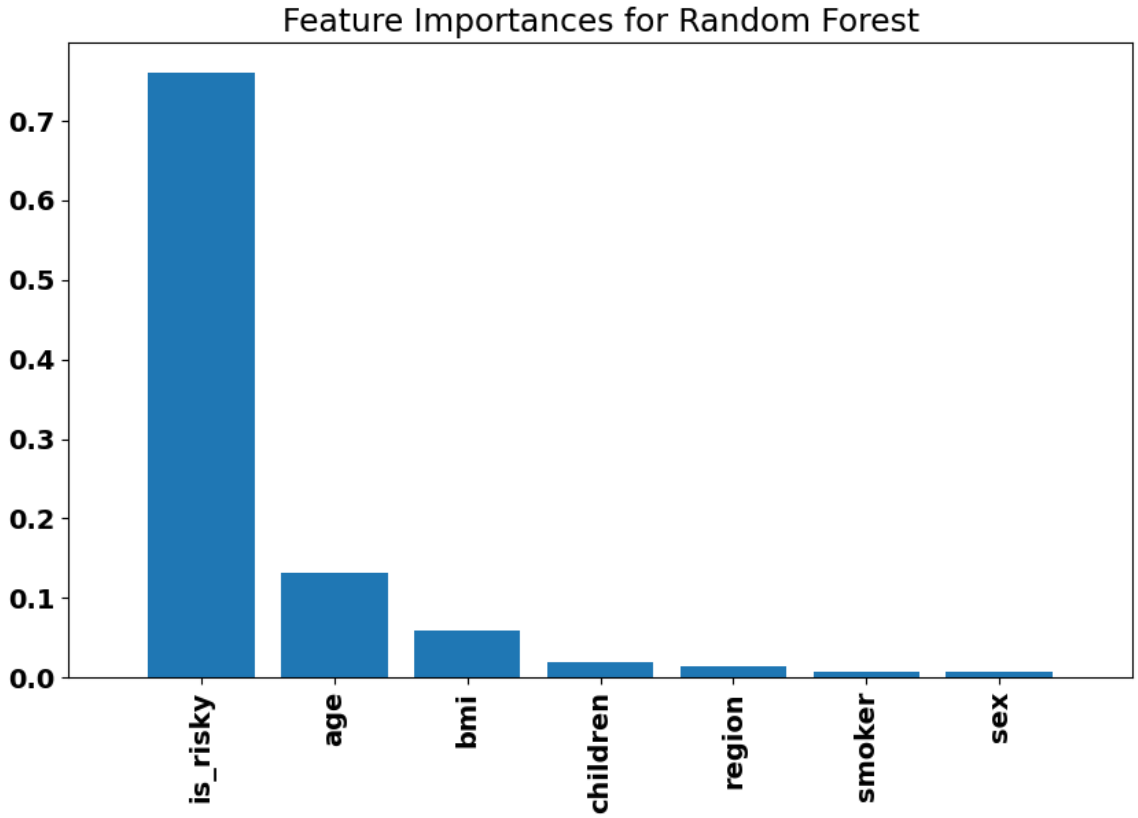


7. Feature Importance

Top Features for Random Forest and Decision Tree:

Top Features for Random Forest and Decision Tree

Feature	Importance
is_risky	Highest
age	Moderate
bmi	Moderate



`is_risky` proved to be the most influential predictor.

8. Hyperparameter Tuning

Grid Search was performed on Random Forest:
- n_estimators: [100, 200, 300]
- max_depth: [5, 10, 15]
- min_samples_split: [2, 5, 10]

Best R^2 after tuning: 0.8473

9. Summary

- Random Forest outperformed other models in accuracy and generalization
- Feature Engineering using `is_risky` had significant impact
- No manual encoding was needed due to auto-cleaning efficiency
- The workflow includes solid validation steps including cross-validation and residual analysis