Insurance Prediction Model

A Comprehensive Analysis Using Machine Learning Algorithms

Objective

1

Analyze factors affecting medical expenses.

2

Compare and evaluate multiple regression and machine learning models.

3

Develop an interactive dashboard for predictions and insights.

Dataset Overview

The dataset contains 1338 rows and 6 columns:

- age: Age of the individual
- sex: Gender (Male, Female)
- bmi: Body Mass Index

- children: Number of children/dependents
- smoker: Whether the person is a smoker (Yes, No)

- region: Geographic region of the individual (Southwest, Southeast, Northwest, Northeast)

- expenses: Yearly medical expenses (target variable)

Removed missing values.

Maintained original and encoded versions for different use cases.

Data Preprocessing

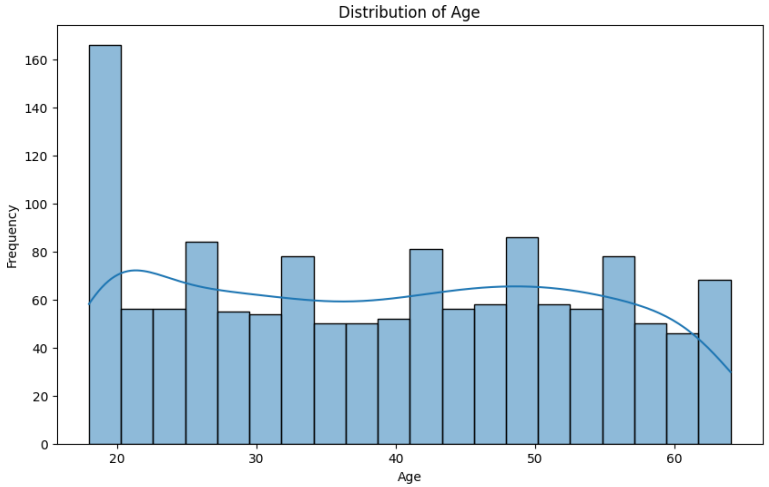
Categorical Encoding: Sex, Smoker, and Region.

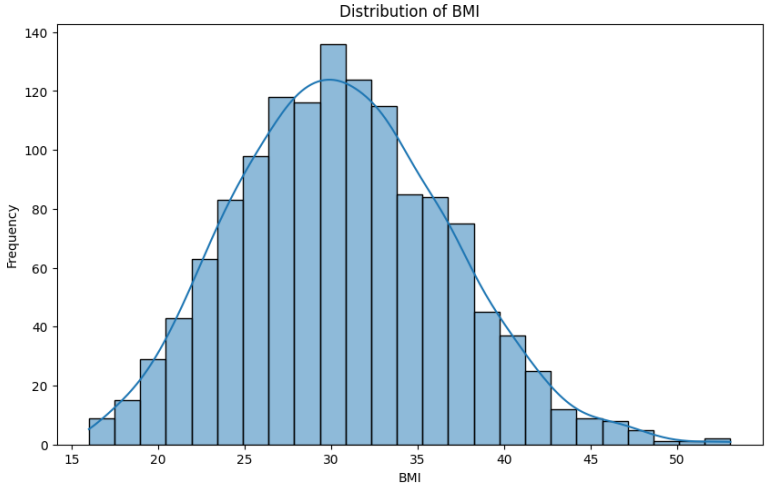
Scaling: Applied MinMaxScaler to Age, BMI, and Children.

Feature Engineering (if applicable): Added interaction terms like BMI-Smoker.

Exploratory Data Analysis (EDA)

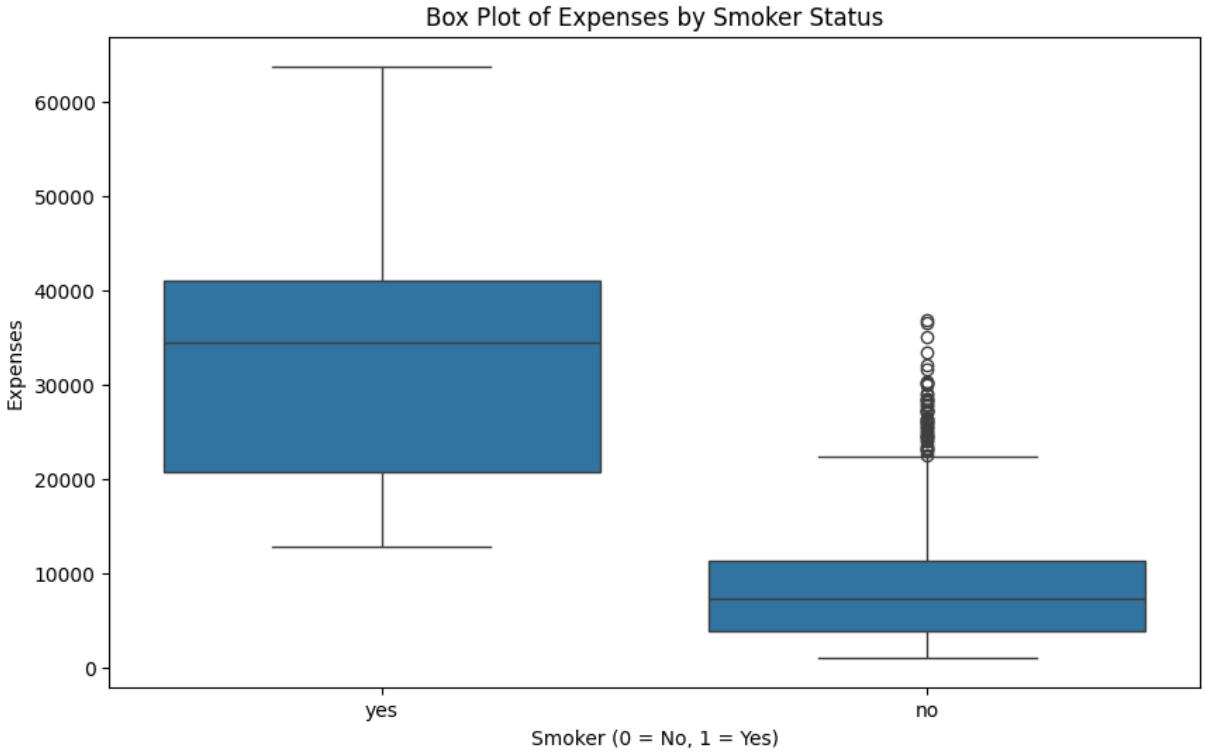
- Strong correlation between 'bmi' and 'expenses'
- Smokers tend to have higher expenses
- Age also correlates with medical expenses, with older individuals incurring more costs



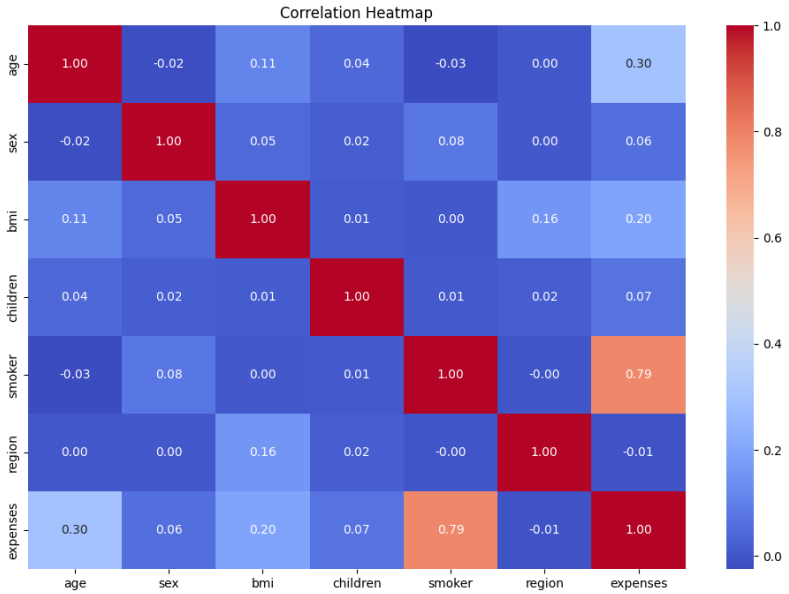


Histograms for distributions of Age and BMI

Boxplots comparing Smoker vs. Expenses



Correlation Heatmap



Model Selection

Linear Regression

Multiple Linear Regression

Decision Tree Regressor

Random Forest Regressor

Support Vector Regression (SVR)

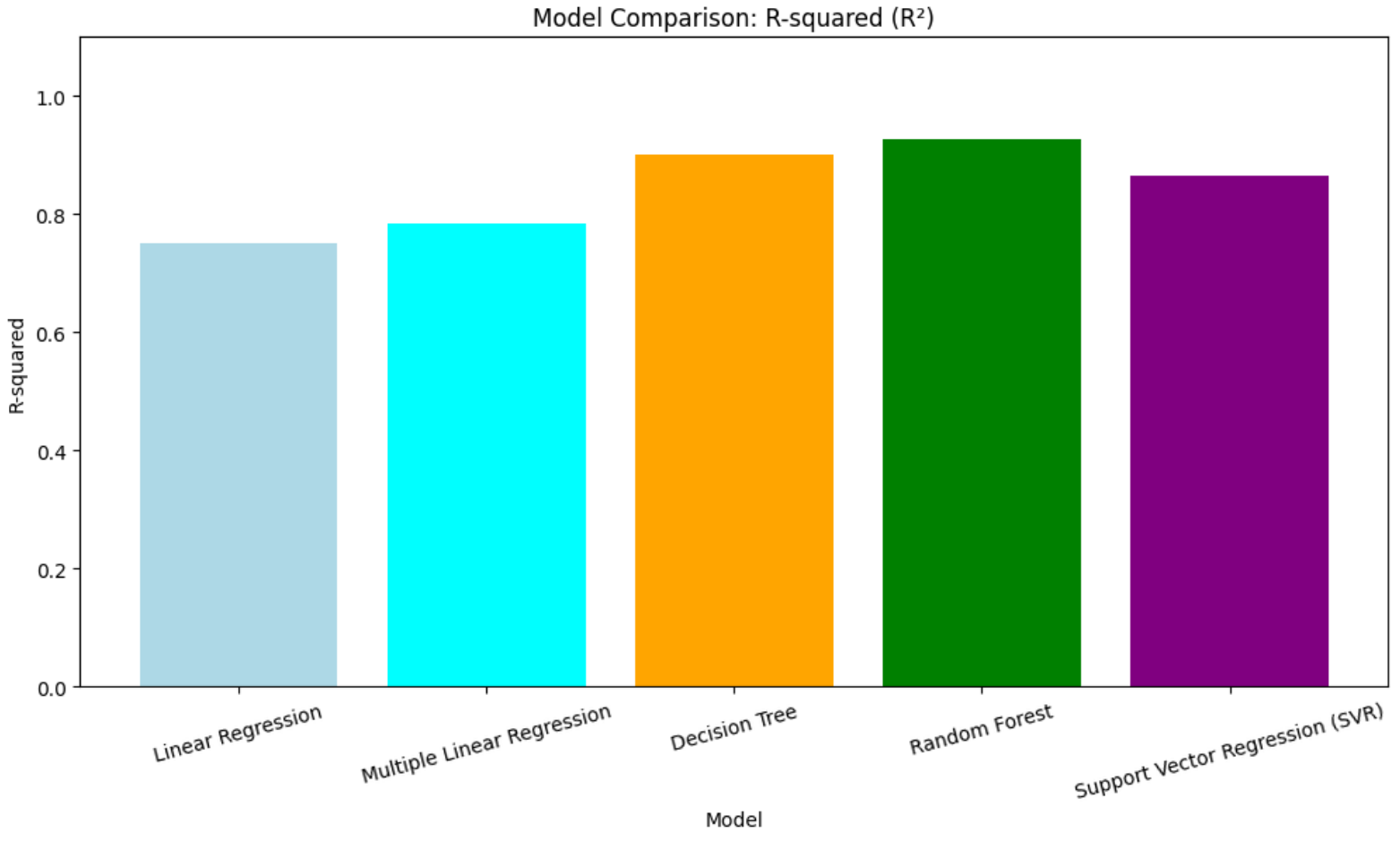
Model Performance

Mean Squared Error (MSE):	Linear Regression: 36,525,540
	Multiple Linear Regression: 33,639,080
	Decision Tree Regressor: 15,425,350
	Random Forest Regressor: 11,175,610
	SVR: 20,956,870
R-squared (R^2):	Linear Regression: 0.750
	Multiple Linear Regression: 0.783
	Decision Tree Regressor: 0.901
	Random Forest Regressor: 0.928
	SVR: 0.865

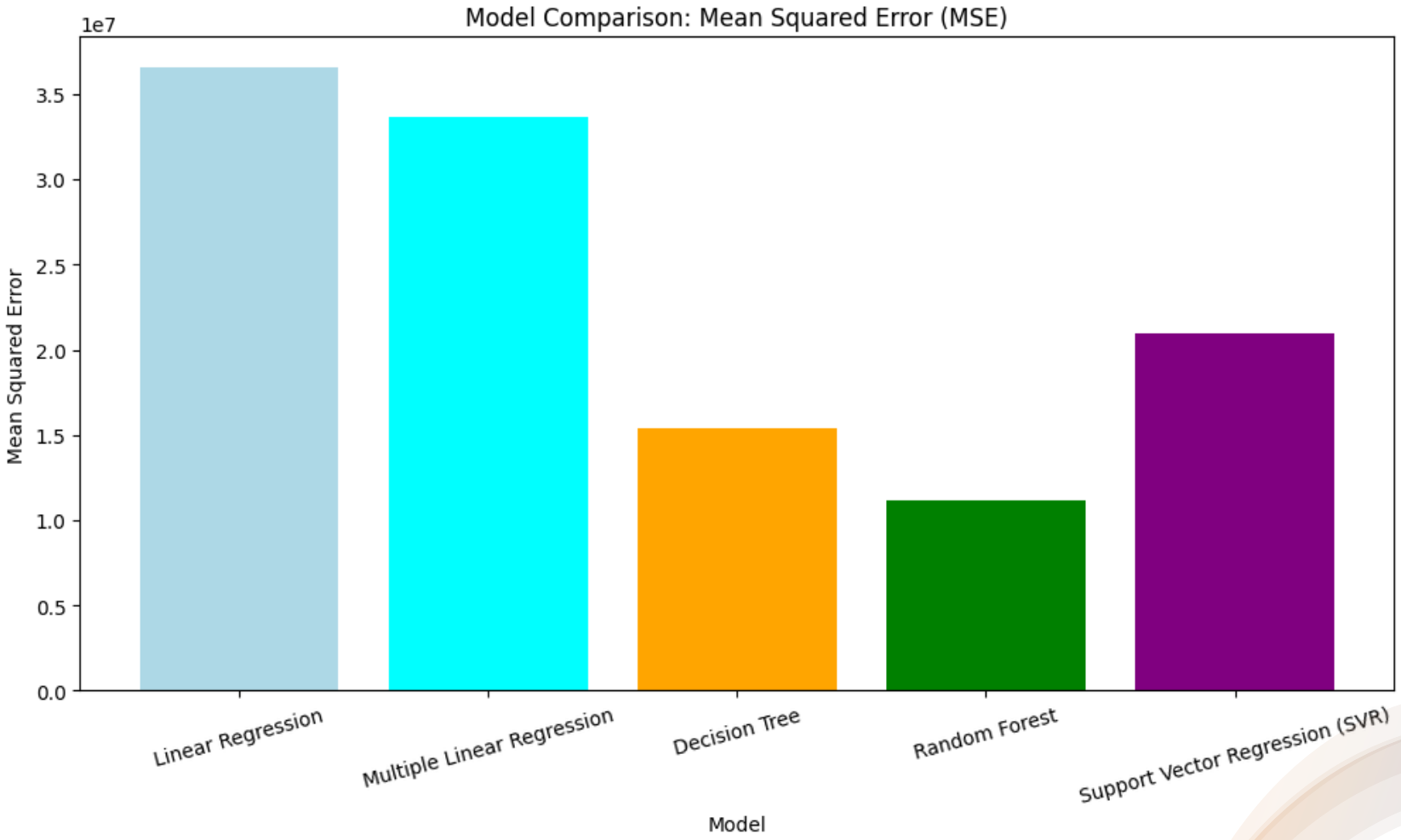
Model Comparison Table:

	Model	Mean Squared Error (MSE)	R-squared (R^2)
0	Linear Regression	3.652554e+07	0.750752
1	Multiple Linear Regression	3.363908e+07	0.783321
2	Decision Tree	1.542535e+07	0.900641
3	Random Forest	1.117561e+07	0.928015
4	Support Vector Regression (SVR)	2.095687e+07	0.865011

Model Performance

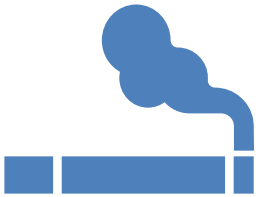


Model Performance



Model Performance

Insights and Conclusions



Key Insights:

Smoking is a significant factor in predicting medical expenses.

Tree-based models handle non-linear relationships effectively.

Random Forest outperforms other models with the lowest MSE and highest R^2.

Dashboard Benefits:

Interactive prediction of medical expenses.

Visual analysis for better understanding of feature relationships.

Thank You!