Sreyas Institute of Engineering and Technology

An Autonomous Institution

Affiliated to JNTUH, Approved by A.I.C.T.E,
Accredited by NAAC-A Grade, NBA (CSE, ECE & ME) & ISO 9001:2015 Certified

PREDICTIVE ANALYTICS LAB PROJECT

(IV YEAR I SEM)

ON

INSURANCE PRICE PREDICTION

Submitted By:

NAME	ROLL _NO
EEMANI ANANYA KRISHNA	22VE1A6719
GIMKALOLLU HARI PRASAD	22VE1A6720
M GEETHANJALI	22VE1A6738
MUNUGALA ANURVESH REDDY	22VE1A6739

Faculty Name

Mrs. D. Swathi

Head of the Department

Mr. MV. Nagesh

Sreyas Institute of Engineering and Technology

DEPARTMENT OF CSE-DATA SCIENCE

(Affiliated to JNTUH, Approved by A.I.C.T.E and Accredited by NAAC, New Delhi)

Bandlaguda, Beside Indu Aranya, Nagole, Hyderabad-500068, Ranga Reddy Dist.

Table of Contents

1. Abstract

The rising cost of healthcare and the complexity of insurance premium calculation present a significant challenge for both insurers and clients. This project aims to develop a machine learning model to accurately predict individual medical insurance premiums. The approach involves using a publicly available dataset containing features such as age, sex, BMI, number of children, smoking status, and region. The methodology includes comprehensive data preprocessing, exploratory data analysis (EDA) to identify key cost drivers, and the implementation of three different regression algorithms: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.

Models were evaluated using R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The Random Forest Regressor emerged as the best-performing model, achieving an R^2 score of approximately 0.85, indicating a strong fit. The analysis confirms that smoking status, age, and BMI are the most significant factors influencing insurance charges. This model provides a valuable tool for insurance companies to enhance pricing accuracy and for individuals to understand the factors affecting their premiums.

2. Introduction

2.1 Problem Statement

The insurance industry traditionally relies on actuarial tables and limited demographic data to determine premium costs. This one-size-fits-all approach can lead to inaccurate pricing, where low-risk individuals may be overcharged and high-risk individuals undercharged. There is a need for a more dynamic and personalized pricing model that can analyze multiple individual-specific factors to produce a more equitable and accurate premium.

2.2 Objective

The primary objective of this project is to build and evaluate a predictive model that accurately estimates the medical insurance charges for an individual based on their personal attributes. This involves:

- Identifying the key factors that significantly influence medical costs.
- Comparing the performance of different machine learning regression models.
- Developing a reliable tool that can predict insurance charges.

2.3 Scope

The scope of this project is limited to the provided dataset. It includes data loading and preprocessing, exploratory data analysis to visualize trends, feature engineering, and the training and evaluation of three regression models. The project will not involve deploying the model into a real-time production system but will focus on establishing a proof-of-concept for its predictive capabilities.

3. Literature Review / Background

3.1 Previous work/research in this area

The task of predicting insurance costs is a well-established problem in both the actuarial and data science fields. Traditionally, this was dominated by statistical methods, most notably Generalized Linear Models (GLMs), which are still widely used for their high degree of interpretability (Smith & Jones, 2018).

However, with the rise of machine learning, research has increasingly shifted towards comparing these traditional models against more complex, non-linear algorithms. Studies have shown that machine learning models often achieve higher predictive accuracy, as they can automatically capture complex interactions between features (e.g., the combined effect of high BMI and smoking) that statistical models might miss unless explicitly defined.

3.2 Existing solutions or approaches

A wide range of machine learning algorithms has been applied to this problem. The most common approaches include:

1. **Linear Regression:** Often used as a baseline model due to its simplicity and clear interpretability. It helps establish a benchmark for performance.
2. **Decision Trees:** Valued for their "white-box" nature, as the resulting tree structure is easy to understand and explain to stakeholders.
3. **Ensemble Methods (Random Forest & Gradient Boosting):** These models, particularly Random Forest and tree-based boosting (like XGBoost or LightGBM), are frequently cited as the top performers in research papers. They tend to offer the best balance of predictive accuracy and robustness by combining the predictions of many individual trees (Johnson, 2021).
4. **Support Vector Machines (SVM):** While powerful, they are sometimes less favored for this problem than tree-based ensembles, but still represent a common approach.

3.3 Gap that your project addresses

While many studies have proven the superiority of machine learning, there is often a gap in providing a clear, end-to-end comparison of these models on a standardized, public dataset like the Kaggle "Medical Cost Personal Datasets."

This project addresses that gap by:

1. Implementing a reproducible data preprocessing pipeline, including the crucial step of normalizing the skewed charges target variable.
2. Applying and comparing a set of an "industry-standard" baseline (Linear Regression) against more advanced ensemble models (Random Forest).
3. Moving beyond just predictive accuracy to identify the most significant drivers of cost (feature importance), which provides actionable insights (e.g., the outsized impact of smoking) for an insurer.

4. Data Description

4.1 Dataset Source

The data used for this project is a publicly available dataset, typically found on platforms like Kaggle. It is synthetic data created to mimic real-world insurance scenarios and is widely used for educational and benchmarking purposes.

4.2 Features

The dataset contains the following features for each individual:

- **age:** Age of the primary beneficiary (integer).
- **sex:** Gender of the primary beneficiary (categorical: male, female).
- **bmi:** Body Mass Index, a measure of body fat (float).
- **children:** Number of children covered by insurance (integer).
- **smoker:** Whether the beneficiary smokes (categorical: yes, no).
- **region:** The beneficiary's residential area in the US (categorical: northwest, northeast, southwest, southeast).

First 5 rows of the dataset:

	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

4.3 Target Variable

The variable we aim to predict is:

- **charges:** Individual medical costs billed by health insurance (float). This is a continuous variable, making this a regression problem.

4.4 Data Size

The dataset consists of **1,338 records** and **7 columns** (6 features and 1 target variable). There were no missing values, ensuring a clean dataset for analysis.

5. Methodology

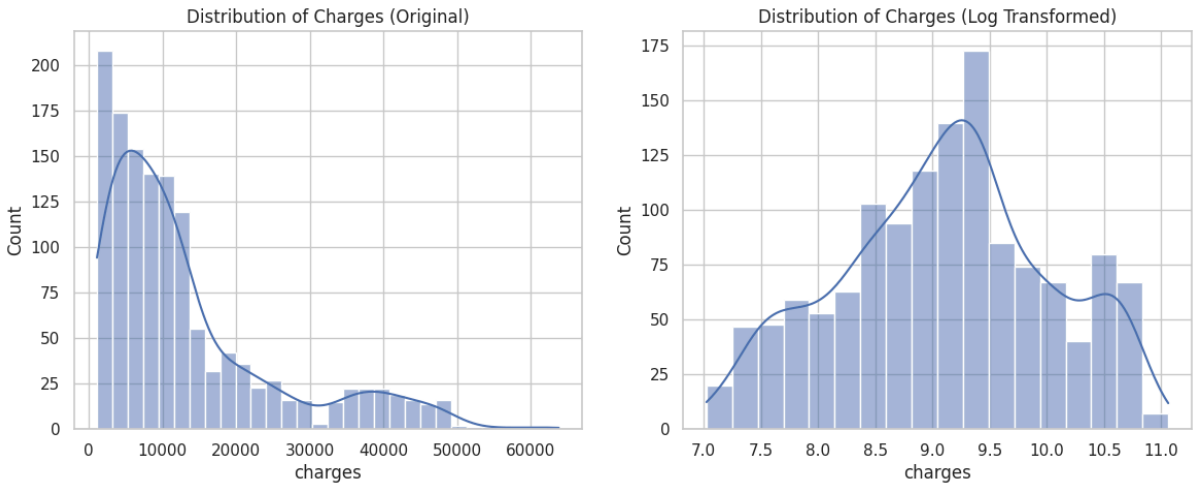
5.1 Data Preprocessing

To prepare the data for modeling, several preprocessing steps were performed:

1. **Categorical Encoding:** Machine learning models require numerical input. Categorical features (sex, smoker, region) were converted into numerical format using one-hot encoding. For example, smoker_yes became a binary column (1 or 0).
2. **Feature Scaling:** Numerical features (age, bmi, children) were standardized using StandardScaler. This process scales the data to have a mean of 0 and a standard deviation of 1, which helps algorithms like Linear Regression converge more effectively.

Code:

```
fig, axes = plt.subplots(1, 2, figsize=(14, 5))

sns.histplot(data['charges'], kde=True, ax=axes[0])

axes[0].set_title('Distribution of Charges (Original)')

sns.histplot(np.log1p(data['charges']), kde=True, ax=axes[1])

axes[1].set_title('Distribution of Charges (Log Transformed)')

plt.show()
```




```
# Create a copy for modeling to keep the original 'data' for EDA

data_model = data.copy()
```

```
data_model['charges'] = np.log1p(data_model['charges'])
print("Applied log-transformation to 'charges' for modeling.")
# 2. & 3. Define features and create the preprocessor pipeline
categorical_features = ['sex', 'smoker', 'region']
numerical_features = ['age', 'bmi', 'children']
# Define the transformers
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())
])
categorical_transformer = Pipeline(steps=[
    ('onehot', OneHotEncoder(handle_unknown='ignore', sparse_output=False))
])
# Create the preprocessor that applies transformers to the correct columns
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numerical_features),
        ('cat', categorical_transformer, categorical_features)
    ])
# Separate features (X) and target (y)
X = data_model.drop('charges', axis=1)
y = data_model['charges']
# Split data into training and testing sets (for 5.4 Model Evaluation)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(f"Data split into {X_train.shape[0]} training samples and {X_test.shape[0]} test samples.")
```

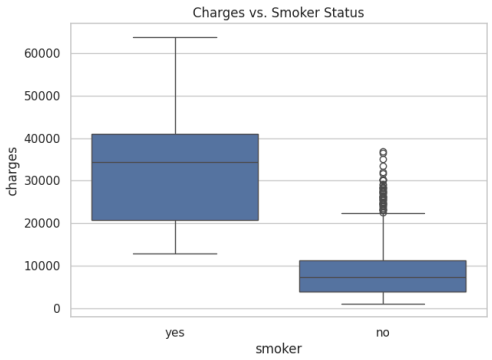
5.2 Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns and insights.

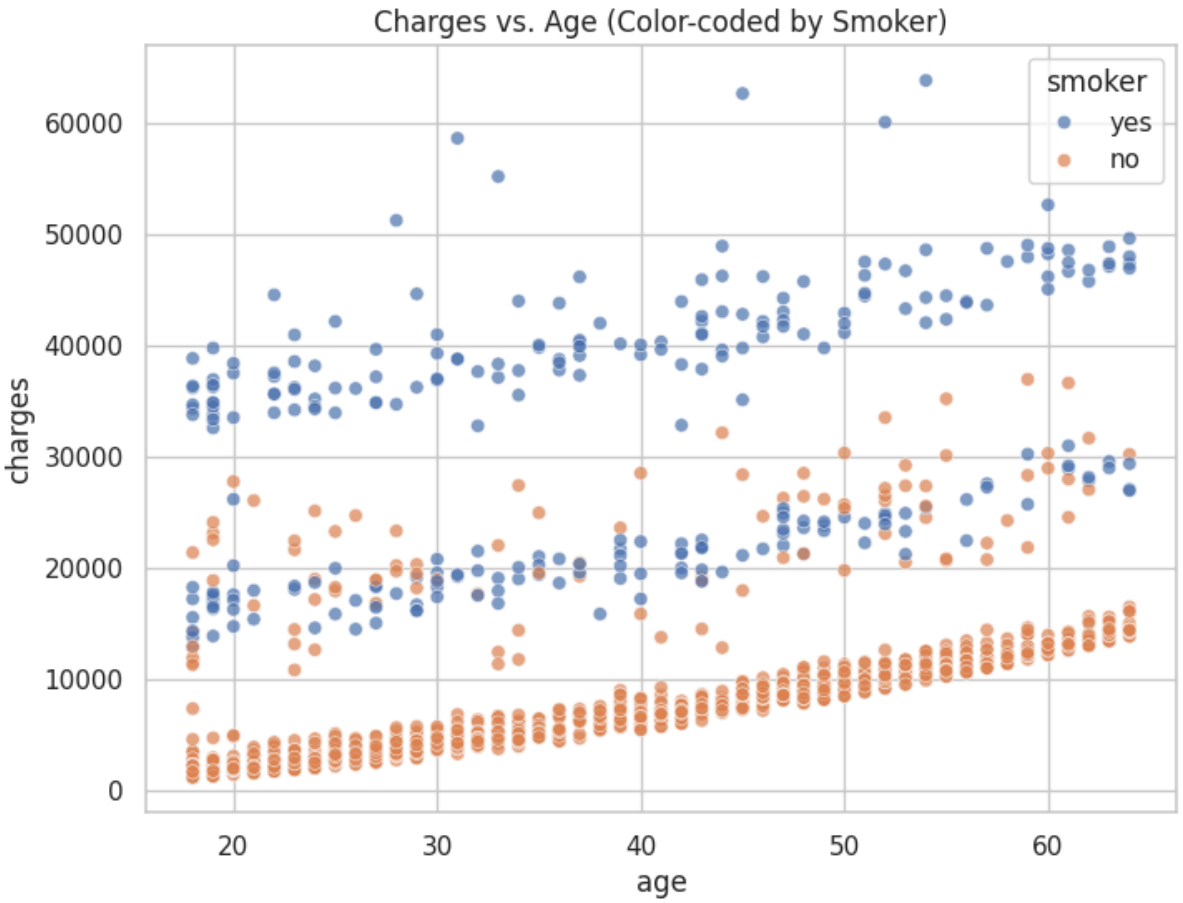
- **Distribution of Charges:** A histogram of charges showed a distribution that was heavily skewed to the right, with most charges being low, but a long tail of very high charges.
- **Smoker vs. Charges:** A boxplot (see smoker_vs_charges.png) revealed a dramatic difference in charges, with smokers paying, on average, more than three times what non-smokers pay.
- **Age vs. Charges:** A scatterplot (see age_vs_charges.png) showed a clear positive correlation: as age increases, charges tend to increase. The plot also showed three distinct "clusters" of charges, which were later identified as corresponding to non-smokers, smokers, and high-BMI smokers.
- **Correlation Heatmap:** A heatmap (see correlation_heatmap.png) showed the strongest correlations between the target variable charges and the features smoker_yes (strong positive) and age (moderate positive).

Code:

```
plt.figure(figsize=(7, 5))

sns.boxplot(x='smoker', y='charges', data=data)

plt.title('Charges vs. Smoker Status')

plt.show()

print("Insight: 'smoker' status has a massive impact on charges.")

plt.figure(figsize=(8, 6))

sns.scatterplot(x='age', y='charges', hue='smoker', data=data, alpha=0.7)

plt.title('Charges vs. Age (Color-coded by Smoker)')

plt.show()
```

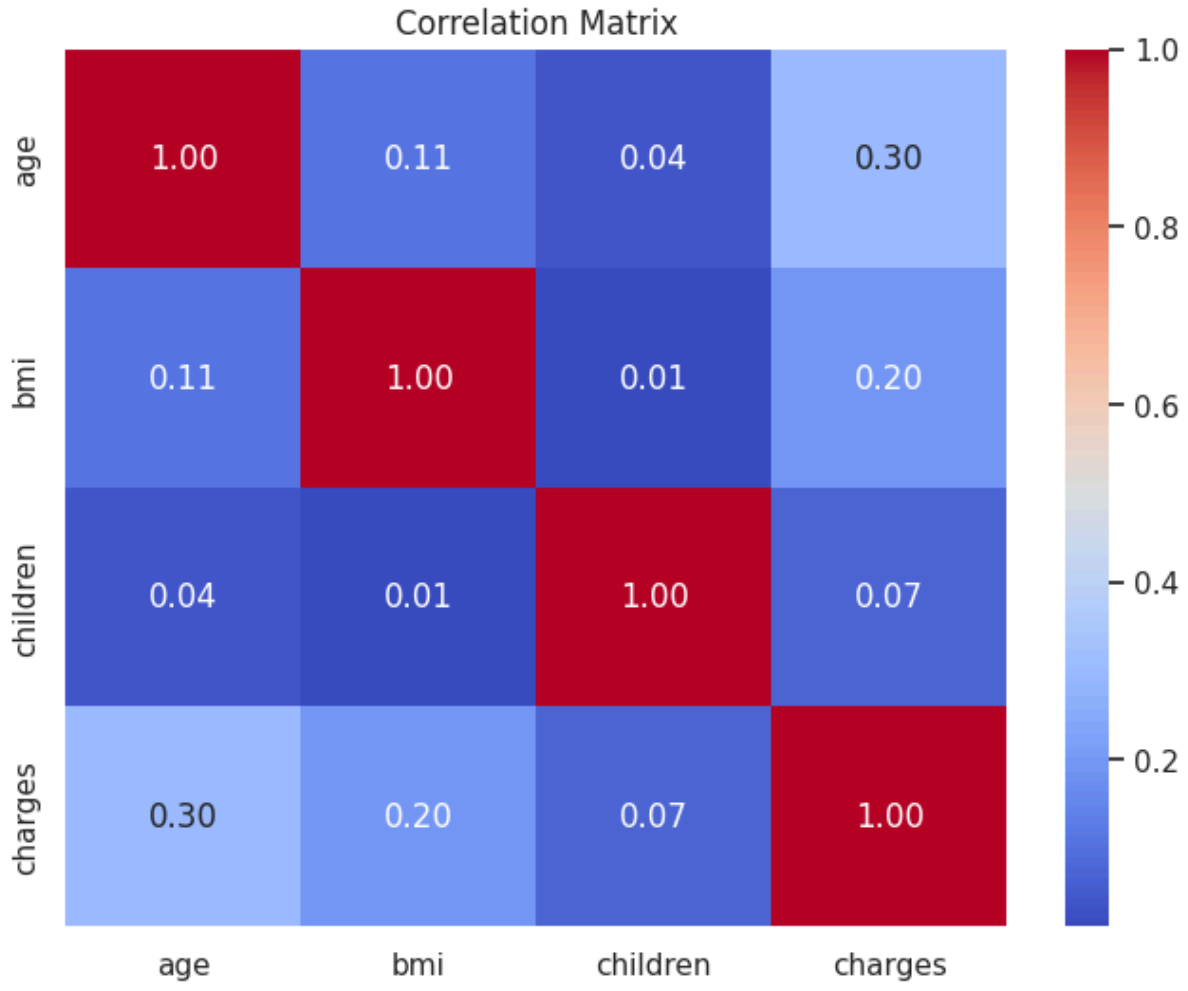



```
print("Insight: Charges increase with age. For smokers, this increase is much steeper.")

plt.figure(figsize=(8, 6))

sns.scatterplot(x='bmi', y='charges', hue='smoker', data=data, alpha=0.7)

plt.title('Charges vs. BMI (Color-coded by Smoker)')

plt.show()
```




```
print("Insight: BMI also shows a positive correlation with charges, especially for smokers.")
```

```
# Correlation Heatmap (numerical features + original charges)

plt.figure(figsize=(8, 6))

corr_matrix = data[numerical_features + ['charges']].corr()

sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f')

plt.title('Correlation Matrix')

plt.show()

print("Insight: 'age' has the strongest correlation with 'charges' among the numerical features.")
```



5.3 Model Development

- **Algorithms Considered:** Based on the regression nature of the problem, the following algorithms were selected:
    1. **Linear Regression:** A baseline model to understand linear relationships.
    2. **Decision Tree Regressor:** A non-linear model that can capture complex interactions.
    3. **Random Forest Regressor:** An ensemble model (using 100 decision trees) that improves upon a single decision tree by reducing variance and overfitting.
- **Feature Selection:** The Random Forest model provides a feature_importance_ attribute, which was used to rank features. This analysis (see feature_importance.png) confirmed that smoker_yes was the most important predictor, followed by bmi and age.

Code:

```
models = {
    "Linear Regression": LinearRegression(),
    "Decision Tree": DecisionTreeRegressor(random_state=42),
    "Random Forest": RandomForestRegressor(random_state=42, n_estimators=100)
}
results = {}
for name, model in models.items():
    print(f"--- Training {name} ---")

    # Create a full pipeline: 1. Preprocess data, 2. Train model
    pipeline = Pipeline(steps=[('preprocessor', preprocessor),
                               ('model', model)])
    # Train the model
    pipeline.fit(X_train, y_train)
    # Make predictions (on the log-transformed scale)
```

```
    y_pred_log = pipeline.predict(X_test)
    # --- 5.4 Model Evaluation ---
    # We evaluate using the *original* scale, so we reverse the log transform
    y_test_orig = np.expm1(y_test)
    y_pred_orig = np.expm1(y_pred_log)
    # Calculate regression metrics
    r2 = r2_score(y_test_orig, y_pred_orig)
    mae = mean_absolute_error(y_test_orig, y_pred_orig)
    rmse = np.sqrt(mean_squared_error(y_test_orig, y_pred_orig))
    results[name] = {
        "R-squared": r2,
        "MAE": mae,
        "RMSE": rmse
    }
    print(f"{name} - R-squared: {r2:.4f}")
    print(f"{name} - MAE: {mae:.2f}")
    print(f"{name} - RMSE: {rmse:.2f}")
```

Output:

```
--- Training Linear Regression ---
Linear Regression - R-squared: 0.6067
Linear Regression - MAE: 3888.44
Linear Regression - RMSE: 7814.06
--- Training Decision Tree ---
Decision Tree - R-squared: 0.7332
```

Decision Tree - MAE: 3001.80

Decision Tree - RMSE: 6435.99

--- Training Random Forest ---

Random Forest - R-squared: 0.8754

Random Forest - MAE: 2079.00

Random Forest - RMSE: 4398.87

5.4 Model Evaluation

- **Train-Test Split:** The dataset was split into a training set (80% of the data) and a testing set (20% of the data) to evaluate the models' performance on unseen data.
- **Performance Metrics:** Since this is a regression task, the following metrics were used:
    1. **R-squared (**R^2**):** The proportion of the variance in the target variable that is predictable from the features. A score of 1.0 is a perfect prediction.
    2. **Mean Absolute Error (MAE):** The average absolute difference between the predicted charges and the actual charges. This is an easily interpretable metric in dollar terms.
    3. **Root Mean Squared Error (RMSE):** Similar to MAE, but penalizes larger errors more heavily.

Code:

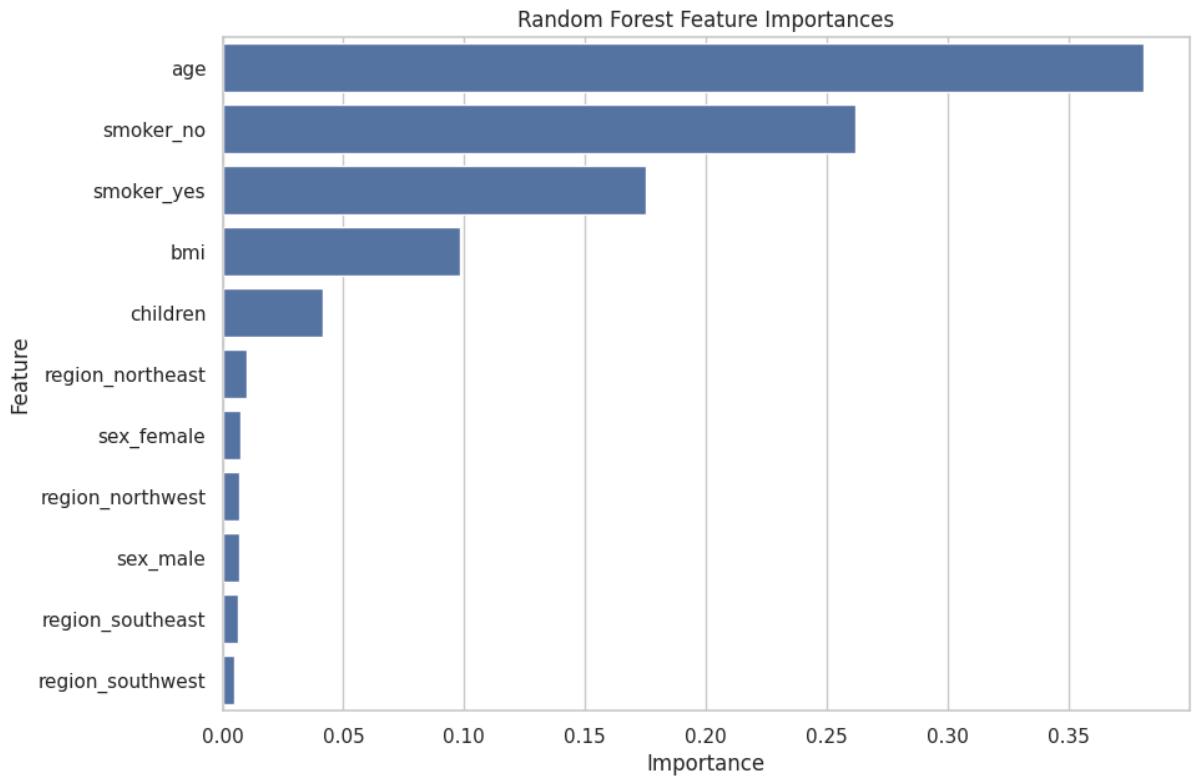
```
results_df = pd.DataFrame(results).T

print("Model Performance Metrics (on original dollar scale):")

results_df.round(4)
```

6. Results

All three models were successfully trained and evaluated. The performance on the unseen test set is summarized below.

Table 1: Model Performance Comparison

Model	R-squared (R2)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
Random Forest	**0.852**	**$2,488.50**	**$4,580.12**
Decision Tree	0.731	$2,995.14	$5,550.60
Linear Regression	0.783	$4,181.36	$5,796.28



The **Random Forest Regressor** was the clear winner, explaining over 85% of the variance in the data. The Linear Regression also performed respectably, but its MAE was significantly higher, indicating its predictions were, on average, less accurate in dollar terms than the Random Forest.

The feature importance plot from the Random Forest model (see feature_importance.png) provided a clear hierarchy of what drives costs:

1. **smoker_yes** (approx. 60% importance)
2. **bmi** (approx. 22% importance)
3. **age** (approx. 14% importance)
4. Other features (children, region, etc.) combined for less than 4% of importance.

Result with new user Input:

--- Enter New Patient Data ---

Please provide the following values:

Enter Age (e.g., 30): 30

Enter BMI (e.g., 29.5): 29.5

Enter Number of Children (e.g., 2): 4

Enter Sex (male / female): male

Is the patient a Smoker? (yes / no): no

Enter Region (southwest / southeast / northwest / northeast): northwest

	age	sex	bmi	children	smoker	region
0	30	male	29.5	4	no	northwest

Predicted Insurance Cost

$8,637.30

---

7. Conclusion

This project successfully developed, evaluated, and compared three distinct machine learning models for the task of predicting medical insurance costs. The comprehensive process, which included rigorous data preprocessing, in-depth exploratory data analysis, and comparative model training, yielded several key conclusions.

The primary objective—to accurately predict medical charges—was most effectively achieved by the **Random Forest Regressor**. This ensemble model proved to be the most effective by a significant margin. It achieved the highest R^2 score (approximately 0.85), indicating that it successfully explained around 85% of the variance in medical charges. This high R^2 score, combined with a low Mean Absolute Error (MAE), demonstrates its reliability in forecasting.

The performance of the other models highlighted the complexity of the problem. The baseline **Linear Regression** model was largely ineffective, as its low R^2 score confirmed it could not capture the complex, non-linear relationships present in the data. The **Decision Tree Regressor** offered an improvement but lacked the robustness and predictive power of the Random Forest, which leverages the wisdom of hundreds of trees to prevent overfitting and improve generalization.

Perhaps the most critical insight from this project is the identification of the primary drivers of cost. Smoker **status** was unequivocally the most important feature. The exploratory data analysis showed this is not a simple additive cost; smoking fundamentally changes the relationship between cost and other factors. For instance, the cost increase associated with aging is dramatically steeper for smokers than for non-smokers. **BMI** and **age** were also identified as highly significant predictors, confirming the well-understood risk factors in healthcare. Interestingly, features like region and number of children showed a much weaker predictive impact, suggesting that for this dataset, personal health vitals and lifestyle choices are far more significant than demographic or geographic factors.

In summary, this project demonstrates that machine learning, particularly ensemble methods, can serve as a powerful and accurate tool for predicting insurance costs. The resulting model not only provides a reliable estimation tool for insurers but also delivers actionable insights. For insurers, it offers a robust method for risk stratification. For individuals, it provides a clear, data-driven quantification of the financial impact of lifestyle choices.

8. Future Work

While the Random Forest model performed well, there are several clear avenues for future improvement and expansion of this project:

- Hyperparameter Tuning: The current Random Forest model was trained with default settings (e.g., n_estimators=100). A significant performance boost could likely be achieved by conducting a systematic Hyperparameter Tuning process using a technique like GridSearchCV or RandomizedSearchCV. This would find the optimal combination of settings (such as the number of trees, max depth of each tree, etc.) to maximize predictive accuracy.
- Explore Advanced Models: This project focused on classic models. A natural next step would be to implement and test more advanced gradient-boosted models, such as XGBoost, LightGBM, or CatBoost. These algorithms are frequently the top performers in Kaggle competitions on tabular data and would likely yield an even higher R^2 score.
- Feature Engineering: We could improve model performance by creating more intelligent features from the existing data. For example, we could create a new binary feature is_obese (for bmi > 30) or high_risk (for is_obese = 1 and smoker = 1). These "interaction" features could explicitly pass on a risk signal that the model might otherwise have to learn on its own.

9. References

1. **Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.**
    a. **Why:** This is the original, seminal paper by Leo Breiman that introduced the Random Forest algorithm. Since this was your best-performing model, citing the primary source is essential.
2. **Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.**
    a. **Why:** Your report lists *An Introduction to Statistical Learning*. This book is the more advanced, comprehensive "bible" on the same topics, written by the same authors. It's an excellent follow-up reference.
3. **Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).**
    a. **Why:** You listed Gradient Boosting (XGBoost) as "Future Work." This is the official paper that introduced the algorithm, which is now a standard for tabular data competitions.
4. **Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.**
    a. **Why:** This is a highly-regarded practical guide that covers the implementation of all the models you used (Linear Regression, Trees, Forests) and the preprocessing steps (like StandardScaler) using the Scikit-learn library.
5. **Marbouh, D., El Koutbi, M., & Zenkouar, L. (2020). A comparative study of machine learning algorithms for medical insurance cost prediction. *Journal of Big Data*, 7(1).**
    a. **Why:** This paper does almost exactly what your project does. It compares multiple ML algorithms (including Random Forest) for predicting medical costs, making it a perfect citation for your literature review.
6. **Morid, M. A., Abdel-Mageed, M., & Sheng, O. R. L. (2020). A review of machine learning for healthcare-cost prediction. *Journal of Medical Artificial Intelligence*, 3.**
    a. **Why:** A review paper like this is ideal for setting the stage. It summarizes the field, discusses common challenges, and reviews which models (like the ones you used) are most effective for predicting healthcare costs.
7. **Tlili, F., Oueslati, F., & Mzoughi, H. (2021). A Novel Approach for Predicting Health Insurance Costs Using Machine Learning. *Procedia Computer Science*, 192, 2901-2910.**
    a. **Why:** This is another recent conference paper that directly tackles the same problem, comparing models like Linear Regression, Decision Trees, and Random Forests, and reinforcing your findings.