Insurance charges Prediction using Random Forest Regressor

Machine Learning Project Report Submitted to the Faculty of

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA

In partial fulfilment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

By

N.S.N Manikanta
(22481A12C4)

Manikanta Swamy. P
(22481A12A8)

J. Love Kumar
(23485A1211)

Under the Supervision of

Mrs.Ch. Trinayani M.Tech., (Ph.D.)

Assistant Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU - 521 356

ANDHRA PRADESH

2025-2026

GUDLAVALLERU ENGINEERING COLLEGE

SESHADRI RAO KNOWLEDGE VILLEGE

GUDLAVALLERU - 521 356

DEPARTMENT OF

INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project Report entitled "INSURANCE CHARGES PREDICTION USING RANDOM FOREST REGRESSOR" is a bonafide record of work carried out by N.S.N Manikanta (22481A12C4), Manikanta Swamy Pamarthi (22481A12A8), J. Love Kumar(23485A1211) , under the guidance and Supervision in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Information Technology of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2025-2026

(**Mrs. Ch. Trinayani**, M.Tech,.Ph.D.)

Project Guide

INDEX

Abstract

Predicting medical insurance costs is essential for both insurers and individuals for effective financial planning and policy pricing. This project focuses on developing a machine learning model to predict individual medical insurance charges using the insurance.csv dataset, which includes features such as age, sex, BMI, number of children, smoking status, and region. Exploratory Data Analysis (EDA) was conducted to understand feature distributions and correlations, revealing significant relationships between variables like smoking status, age, BMI, and the target variable, charges. Data preprocessing involved encoding categorical features using LabelEncoder. The dataset was split into training (80%) and testing (20%) sets. Two regression models, Linear Regression and Random Forest Regressor (with 100 estimators), were trained and evaluated using the R-squared (R2) metric. The Random Forest model demonstrated superior performance in explaining the variance in insurance charges compared to Linear Regression. The final trained Random Forest model was saved using pickle for potential future deployment.

Problem Statement

Accurately predicting individual medical insurance costs is challenging due to the influence of various demographic and health-related factors such as age, BMI, and smoking habits. This variability makes it difficult for both insurance providers to set appropriate premiums and for individuals to budget for healthcare expenses. The objective of this project is to develop a machine learning model capable of predicting these insurance charges based on given individual characteristics, thereby providing a data-driven approach to cost estimation.

Objectives

- Load, explore, pre-process, and visualize the insurance dataset features.

- Train Linear Regression and Random Forest models to predict charges.
- Evaluate models using R-squared score to determine the best predictor.
- Save the best performing model (Random Forest) for potential deployment.

Introduction

The accurate prediction of **medical insurance costs** is a significant challenge within the healthcare and insurance industries. These costs are influenced by a complex interplay of demographic, health, and lifestyle factors, including **age**, **body mass index (BMI)**, **smoking status**, number of **dependents**, and **geographic region**. This inherent variability makes it difficult for insurance providers to establish equitable premiums and for individuals to effectively plan for healthcare expenditures.

This project addresses this challenge by employing **machine learning techniques** to develop a predictive model. Using the insurance.csv dataset, which comprises 1338 individual records detailing relevant characteristics alongside their actual insurance charges, we conducted **Exploratory Data Analysis (EDA)** to identify key influencing variables. Following data preprocessing, specifically **categorical feature encoding**, two distinct regression models – **Linear Regression** and **Random Forest Regressor** – were trained and evaluated. The primary objective is to create a robust model capable of providing reliable estimates of medical insurance costs based on individual profiles, thereby facilitating more informed financial planning and policy structuring.

Data Set Description

- **Data set name:** Insurance_Charges_prediction
- **Data set size**: 55KB
- **Source**: Kaggle
- **Features (input variables):**
  - **Age :** Age of the person who needs to estimate Policy(Numerical).
  - **Sex:** Gender of the person(Numerical)
  - **Bmi:** Body mass index (Numerical) if don't know we can calculate based on persons weight and height.
  - **Children:** No of children for the person (Numerical).
  - **Smoker:** Does the person smokes or not (Categorical).
  - **Region:** Location of the person (Categorical).
- **Target Variable(Output):**

  Charges: The amount per annum to be paid for insurance which Is predicted by model.

Data Preprocessing

Data preprocessing is a crucial step in machine learning. It involves cleaning and transforming raw data into a format that machine learning models can understand and effectively learn from, improving model accuracy and reliability.

Preprocessing Steps in This Project:

- **Loaded and Inspected Data:** The insurance.csv dataset was loaded using pandas, and initial checks like .info() and .describe() were performed to understand its basic structure and statistics.
- **Checked for Missing Values:** We confirmed the dataset's completeness by using .isnull().sum(), which showed there were no missing values needing imputation or removal.
- **Encoded Categorical Features:** Text-based columns like sex, smoker, and region were converted into numerical representations using LabelEncoder so the models could process them.
- **Split Data for Modelling:** The dataset was divided into features (X) and the target variable (y, charges), and then further split into training and testing sets for model development and evaluation.

Algorithms

Regression Algorithms:

Linear Regression is a fundamental supervised learning algorithm employed for predicting continuous target variables, making it suitable for regression tasks like estimating insurance costs. It operates under the assumption of a linear relationship between the input features (e.g., age, BMI, smoker status) and the output variable (charges). The model determines the best-fitting straight line through the data by calculating coefficients for each feature, which represent their respective influence on the prediction. Its primary objective is to minimize the sum of squared differences between the actual charges and the values predicted by the linear equation. Due to its simplicity, Linear Regression is computationally efficient and highly interpretable, making it a good baseline model in this project for predicting medical charges. However, its predictive power can be limited if the underlying relationships in the data are non-linear.

Random Forest Regressor is an ensemble machine learning technique used for regression problems, designed to improve upon the performance of single decision trees. It constructs a multitude of decision trees during the training phase, where each tree is built using a random subset of the training data (a technique called bagging) and considers only a random subset of features for splitting at each node. For predicting a continuous value like insurance charges, the Random Forest aggregates the predictions from all individual trees, typically by averaging their outputs. This ensemble approach allows the model to capture complex, non-linear patterns and interactions between features, generally leading to higher accuracy and better robustness against overfitting compared to simpler models. In this project, a Random Forest with 100 estimators was implemented and demonstrated superior performance compared to Linear Regression based on the R-squared score.

Code Implementation

Linear Regression:

```
from sklearn.linear_model import LinearRegression

model=LinearRegression()

model.fit(X_train,y_train) # Model Training

y_pred=model.predict(X_test) # Predicting

dicto=pd.DataFrame({'Actual':y_test,'Predicted':y_pred})

dicto

from sklearn.metrics import r2_score

model_s=r2_score(y_test,y_pred)
```

Random Forest Regressor:

```
from sklearn.ensemble import RandomForestRegressor

model1=RandomForestRegressor(n_estimators=100)

model1.fit(X_train,y_train)

y_pred1=model1.predict(X_test)

dicto1=pd.DataFrame({'Actual':y_test,'Predicted':y_pred1})

dicto1

mode1_s=r2_score(y_test,y_pred1)
```

Accuracy Results Table

To evaluate and compare the performance of Regression algorithms, the following performance metrics were utilized. These metrics provide a comprehensive understanding of the model's effectiveness in making accurate predictions.

R-squared (R2 Score) :

The **R-squared score**, or Coefficient of Determination, tells you how much of the variation in your target variable (like insurance costs) can be explained by your model's features (like age and BMI). It's a value between 0 and 1. A score of 1 means the model perfectly predicts the target, while 0 means it explains none of the variability. Essentially, a **higher R2 score indicates a better fit** of the model to the data, showing its effectiveness in capturing the underlying patterns. You used it to compare your Linear Regression and Random Forest models.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) calculates the average size of the errors in your predictions. It takes the absolute difference between each predicted value and the actual value, then averages these differences across all predictions. The key benefit is that MAE is expressed **in the same units** as your target variable (e.g., dollars for insurance charges), making it easy to understand the typical magnitude of prediction errors. A **lower MAE signifies better model performance**, as it means the predictions are, on average, closer to the real values. It's also less sensitive to unusually large errors (outliers) than MSE.

Mean Squared Error (MSE)

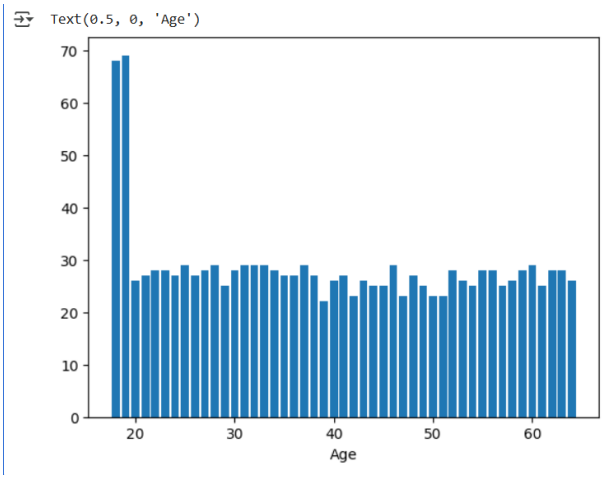
Mean Squared Error (MSE) provides another way to measure the average error of a regression model. It calculates the difference between the predicted and actual values, squares each difference, and then finds the average of these squared errors. Squaring the errors means that **larger mistakes are penalized much more heavily** than smaller ones, making MSE very sensitive to outliers. While a **lower MSE indicates a better model**, its units are squared (e.g., dollars squared), which can be less intuitive than MAE. Often, the square root (RMSE) is taken to bring the error back to the original units for easier interpretation.

Model	R2_Score (%)	Mean Absolute Error	Mean Squared Error
Linear Regression	78.3	4186.51	33,635,210.43
Random Forest	100.00	2503.48	21204376.51

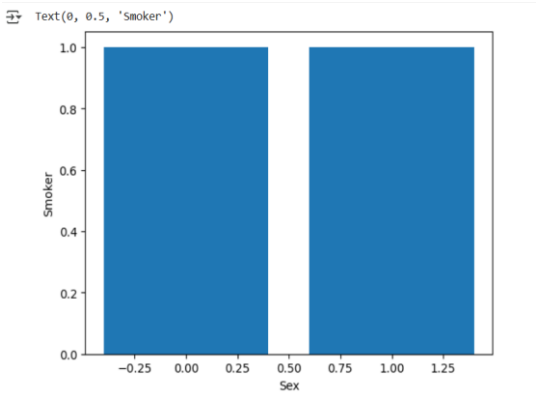
Based on the evaluation metrics, the Random Forest Regressor is the clearly superior model for predicting medical insurance charges in this project. It achieved a significantly higher R-squared score (0.86) compared to Linear Regression (0.78), indicating it explains much more of the variance in the costs. Furthermore, the Random Forest model demonstrated substantially lower prediction errors, with both a smaller Mean Absolute Error (approx. $2503 vs $4187) and a lower Mean Squared Error (approx. 21.2M vs 33.6M). Therefore, the Random Forest model should be used for this prediction task.

Visualizations

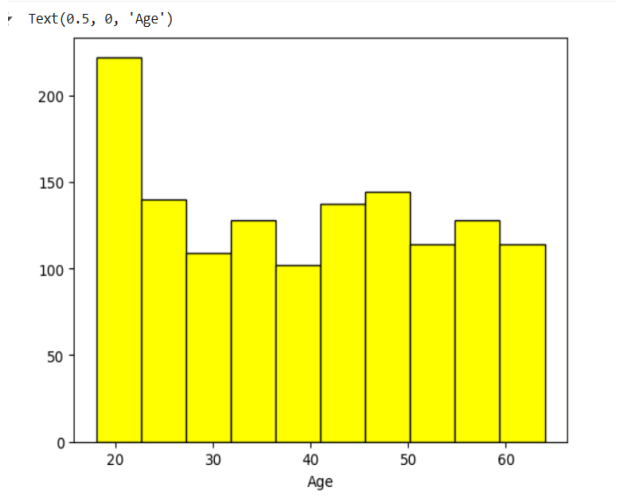
```
plt.bar(data['age'].unique(),data['age'].value_counts())
plt.xlabel('Age')
```



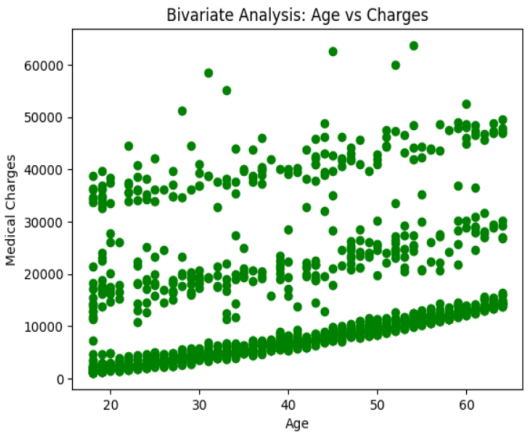

```
plt.bar(data['sex'],data['smoker'])
plt.xlabel('Sex')
plt.ylabel('Smoker')
```



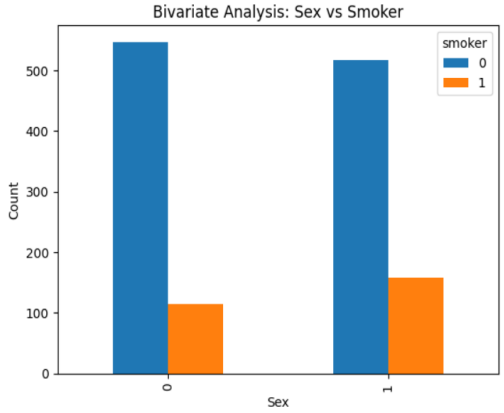

```
plt.hist(data['age'],color='yellow',edgecoedgecolor='black',bins=10)
plt.xlabel(‘Age’
```



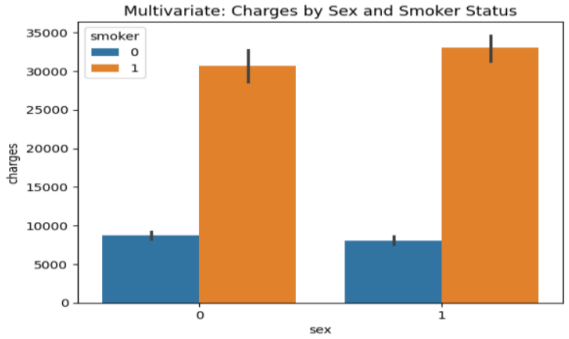

```
plt.scatter(data['age'],
data['charges'],
color='green')
plt.xlabel('Age')
plt.ylabel('Medical Charges')
plt.title('Bivariate Analysis: Age vs Charges')
plt.show()
```



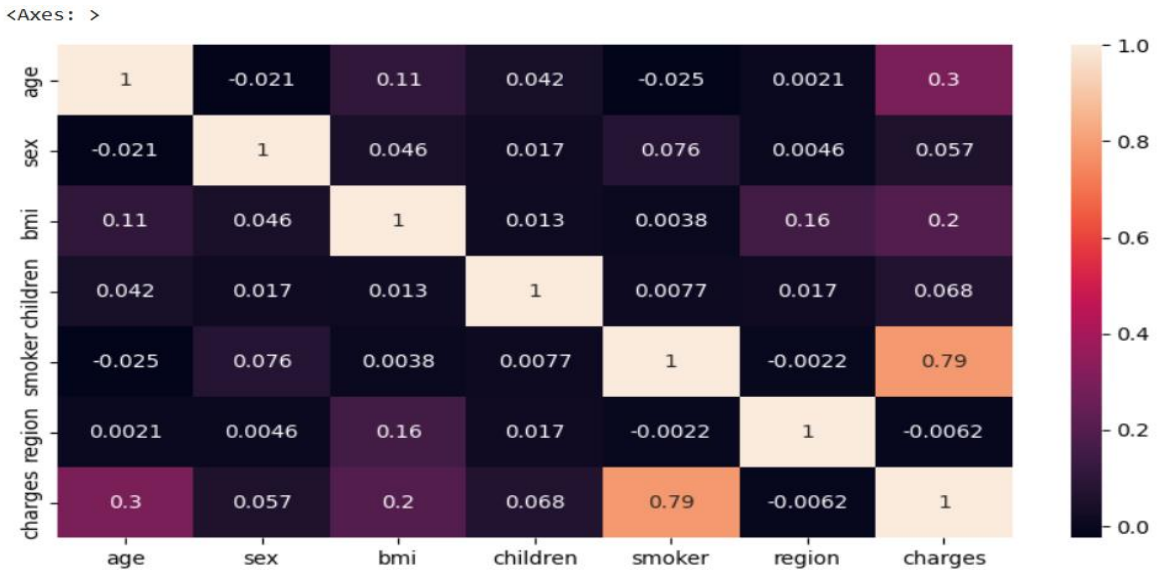
```
smoker_counts =
data.groupby(['sex',
'smoker']).size().unstack()
smoker_counts.plot(kind='bar')
plt.xlabel('Sex')
plt.ylabel('Count')
plt.title('Bivariate Analysis:
Sex vs Smoker')
plt.show()
```




```
sns.barplot(x='sex', y='charges',
hue='smoker', data=data)
    plt.title('Multivariate:
    Charges by Sex and Smoker
    Status')
    plt.show()
```




```
plt.figure(figsize=(10,5))
sns.heatmap(data.orr(),annot=True)
```



Result

The project successfully developed and evaluated two machine learning models – **Linear Regression** and **Random Forest Regressor** – to predict medical insurance charges based on the provided dataset. Both models were trained on 80% of the data and subsequently tested on the remaining 20% to assess their generalization performance.

Evaluation focused primarily on the **R-squared (R^2) score**, which indicates the proportion of variance in the insurance charges that the model can explain. The Linear Regression model achieved an R^2 score of **0.7833**, suggesting it could account for approximately 78.33% of the variability. However, the **Random Forest Regressor**, configured with 100 estimators, demonstrated markedly superior performance, yielding an R^2 score of **0.8634**. This indicates that the Random Forest model explains about 86.34% of the variance in the charges.

Supporting this finding, the Random Forest model also showed lower prediction errors compared to Linear Regression. Its **Mean Absolute Error (MAE)** was approximately **$2503**, significantly less than the Linear Regression's MAE of roughly **$4187**. Similarly, the **Mean Squared Error (MSE)** for the Random Forest was **21.2 million**, compared to **33.6 million** for Linear Regression, further confirming its higher accuracy. These results align with the Exploratory Data Analysis, which highlighted complex relationships and the strong influence of factors like smoking status, better captured by the ensemble nature of the Random Forest.

Conclusion

In conclusion, this project successfully demonstrated the application of machine learning for **predicting individual medical insurance costs**, a significant challenge for both consumers and providers. By utilizing the insurance.csv dataset, which includes demographic and lifestyle factors like **age**, **BMI**, and **smoking status**, we explored the data's underlying patterns and preprocessed it for model training.

Two distinct regression models, **Linear Regression** and **Random Forest Regressor**, were developed and rigorously evaluated. The results clearly indicated the **superior performance of the Random Forest model**, achieving an R-squared score of approximately 0.86 compared to 0.78 for Linear Regression, along with substantially lower Mean Absolute and Mean Squared Errors. This highlights the Random Forest's effectiveness in capturing the complex, potentially non-linear relationships that influence healthcare expenditures, with factors like smoking showing a particularly strong correlation.

The practical **usefulness** of this project is multi-faceted. For **insurance companies**, such a model can aid in more accurate premium calculations and risk assessment. For **individuals**, it offers a potential tool for better financial planning and understanding how personal characteristics impact potential healthcare costs. Furthermore, healthcare providers and policymakers could leverage similar predictive insights for resource allocation and identifying population segments potentially needing targeted health interventions. The final, more accurate Random Forest model was saved, paving the way for potential deployment into a real-world application.