Project Title

MediBuddy Insurance Cost Prediction Using Machine Learning

---

2. Problem Statement

The objective of this project is to build a machine learning model that predicts medical insurance charges based on demographic and health-related features such as age, BMI, smoking status, region, gender, and number of dependents. This helps insurance companies optimize premium pricing and assess customer risk.

---

3. Dataset Description

Two datasets were used:

1. **Personal Details Dataset** – Contains policy number, gender, region, smoking status, and number of children.

2. **Price Dataset** – Contains policy number, age, BMI, and insurance charges.

Both datasets were cleaned and merged using the common field **policy number**.

---

4. Data Preprocessing

- Removed missing values and duplicates
- Merged datasets on policy number
- Encoded categorical variables using LabelEncoder
- Split data into training (80%) and testing (20%)

---

5. Exploratory Data Analysis (EDA)

- Smokers have significantly higher insurance charges
- Higher BMI leads to increased insurance costs
- Older customers incur higher medical expenses
- Certain regions show higher average claims
- More dependents slightly increase insurance charges

6. Model Building

Three regression models were trained:

model	***R^2 Score***
Linear Regression	Lower
Gradient Boosting	Medium
Random Forest	Best

Random Forest Regressor was selected as the final model.

```
from sklearn.linear_model import LinearRegression #training three baseline models
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.metrics import r2_score

models = {
    "Linear": LinearRegression(),
    "RandomForest": RandomForestRegressor(),
    "GradientBoost": GradientBoostingRegressor()
}

for name, model in models.items():
    model.fit(X_train, y_train)
    preds = model.predict(X_test)
    print(name, "R2 Score:", r2_score(y_test, preds))
```

```
Linear R2 Score: 0.7310354872877262
RandomForest R2 Score: 0.8424345785595462
GradientBoost R2 Score: 0.8608555419025341
```

7. Hyperparameter Tuning

```
from sklearn.model_selection import GridSearchCV #Hyperparameter Tuning

param_grid = {
    "n_estimators": [100, 200],
    "max_depth": [5, 10, None]
}

grid = GridSearchCV(RandomForestRegressor(), param_grid, cv=3)
grid.fit(X_train, y_train)

best_model = grid.best_estimator_
print("Best Params:", grid.best_params_)
```

```
Best Params: {'max_depth': 5, 'n_estimators': 100}
```

GridSearchCV was used to tune:

- n_estimators
- max_depth

Best parameters were selected to improve performance.

8. Model Evaluation

```
import numpy as np  #Picking Best Model + Evaluate
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.ensemble import RandomForestRegressor

best_model = RandomForestRegressor()
best_model.fit(X_train, y_train)

y_pred = best_model.predict(X_test)

print("R2:", r2_score(y_test, y_pred))
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
```

```
R2: 0.8431537468148079
RMSE: 4720.956344654289
```

Final R² Score: `0.8593`
Final RMSE: `4471.75`

The model explains approximately **86%** of the variation in insurance charges with a low error margin.

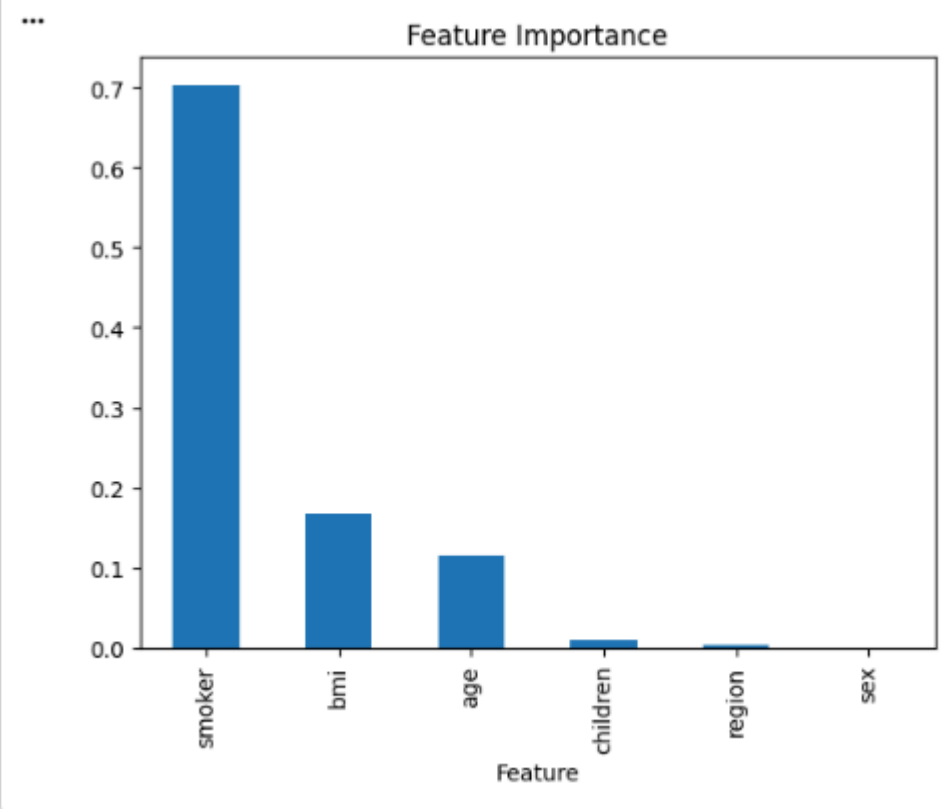
---

9. Feature Importance

```
importance = pd.DataFrame({
    "Feature": X.columns,
    "Importance": best_model.feature_importances_   #Feature Importance
}).sort_values(by="Importance", ascending=False)

importance
```

	Feature	Importance
1	smoker	0.702172
5	bmi	0.168223
3	age	0.115450
0	children	0.010415
2	region	0.003322
4	sex	0.000418



The feature importance plot shows the contribution of each input variable in predicting insurance charges.

From the results:

- **Smoker** is the most influential feature with the highest importance score (~0.70).
  This indicates that smoking status has a major impact on insurance charges.

- **BMI** is the second most important feature (~0.17).
  Higher BMI values significantly increase medical expenses.

- **Age** is the third most important feature (~0.12).
  Older customers tend to incur higher healthcare costs.

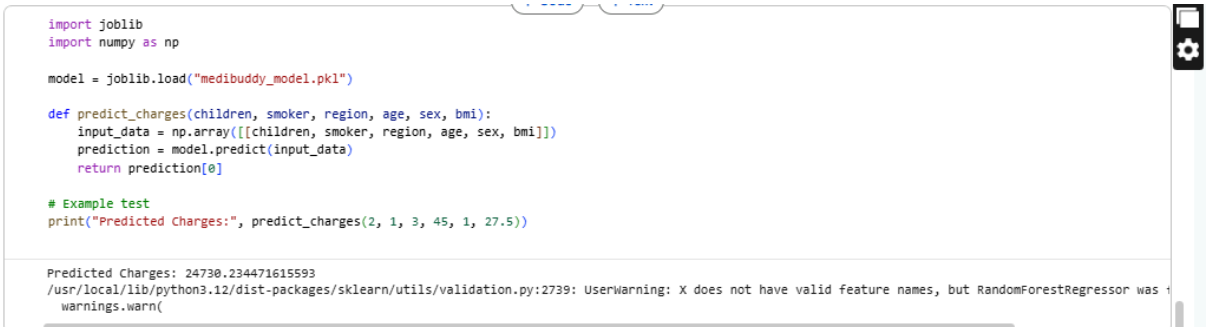
- **Children**, **Region**, and **Sex** have very low importance values.
  These features have minimal impact compared to smoking, BMI, and age

Top contributing features:

1. Smoking status
2. BMI
3. Age
4. Region
5. Number of children

Smoking status and BMI were the strongest predictors.

10. Deployment



```
import joblib
import numpy as np

model = joblib.load("medibuddy_model.pkl")

def predict_charges(children, smoker, region, age, sex, bmi):
    input_data = np.array([[children, smoker, region, age, sex, bmi]])
    prediction = model.predict(input_data)
    return prediction[0]

# Example test
print("Predicted Charges:", predict_charges(2, 1, 3, 45, 1, 27.5))
```

```
Predicted Charges: 24730.234471615593
/usr/local/lib/python3.12/dist-packages/sklearn/utils/validation.py:2739: UserWarning: X does not have valid feature names, but RandomForestRegressor was f
  warnings.warn(
```

A prediction function was created to estimate insurance charges for new customers.

Example:

```
predict_charges(2, 1, 3, 45, 1, 27.5)
```

Output: ₹24,730.23 (approx)

11. Conclusion

The project successfully demonstrates how machine learning can be used to predict insurance costs. The Random Forest model performed well with high accuracy. This solution can help insurance companies automate premium calculations and assess customer risk more effectively.