**📑 Project Report: Healthcare Insurance Premium Prediction**

**1. Abstract**

Healthcare costs vary significantly across individuals depending on age, lifestyle, BMI, and habits such as smoking. Predicting insurance charges accurately helps insurance companies design fair premiums and identify high-risk customers.  
In this project, we developed a Machine Learning model using the **Insurance dataset** to predict medical charges based on demographic and health features. Multiple regression algorithms were compared, and the best-performing model was selected for deployment.

**2. Problem Statement**

The task is to **predict insurance charges** of individuals based on their demographic and medical attributes.

* Business Need: Insurance firms want to **predict future claims** and adjust premium strategies.
* ML Goal: Build a regression model that minimizes prediction error (MAE, RMSE) and maximizes accuracy (R²).

**3. Dataset Description**

* **Source**: Kaggle – Medical Insurance Dataset (insurance.csv)
* **Records**: 1338 samples
* **Features**:

| **Feature** | **Type** | **Description** |
| --- | --- | --- |
| age | Numeric | Age of the individual |
| sex | Categorical | Male/Female |
| bmi | Numeric | Body Mass Index |
| children | Numeric | Number of dependents |
| smoker | Categorical | Smoker/Non-smoker |
| region | Categorical | Region (northeast, northwest, southeast, southwest) |
| charges | Numeric | Insurance charges (Target Variable) |

**4. Data Preprocessing**

1. **Data Inspection**
   * No missing values
   * No duplicate rows
   * Corrected data types
2. **Encoding**
   * sex → Binary encoding (0 = Male, 1 = Female)
   * smoker → Binary encoding (0 = No, 1 = Yes)
   * region → One-Hot Encoding
3. **Feature Scaling**
   * StandardScaler applied to numerical features for models sensitive to scale (Linear Regression, Gradient Boosting).
4. **Train-Test Split**
   * 80% training, 20% testing

**5. Exploratory Data Analysis (EDA)**

* **Univariate Analysis**:
  + Age: Normally distributed
  + BMI: Slight right skew, few high outliers
  + Charges: Right skew, especially high for smokers
* **Bivariate Analysis**:
  + Smokers pay significantly higher charges than non-smokers
  + Higher BMI correlates with higher charges
* **Multivariate Analysis**:
  + Combination of **age + smoker + BMI** has the largest effect on charges

**6. Models Used**

1. **Linear Regression**
2. **Decision Tree Regressor**
3. **Random Forest Regressor**
4. **Gradient Boosting Regressor**

**7. Model Evaluation Metrics**

* **MAE**: Mean Absolute Error → measures average error
* **RMSE**: Root Mean Squared Error → penalizes large errors
* **R² Score**: Goodness of fit → closer to 1 is better

| **Model** | **MAE** | **RMSE** | **R²** |
| --- | --- | --- | --- |
| Linear Regression | High | High | ~0.74 |
| Decision Tree | Moderate | Moderate | ~0.84 |
| Random Forest | Low | Low | ~0.87 |
| Gradient Boosting | Lowest | Lowest | ~0.89 |

**Best Model → Gradient Boosting Regressor** (highest R², lowest RMSE).

**8. Residual Analysis**

* Residuals mostly centered around 0 → model fits well
* Few outliers at very high charges (expected due to rare, extreme smoker cases)

**9. Feature Importance (Random Forest Example)**

* **Smoker** → Most significant feature
* **BMI** → Second highest influence
* **Age** → Strong impact
* **Region & Children** → Minimal effect

**10. Hyperparameter Tuning**

* Applied **GridSearchCV** for Random Forest and Gradient Boosting
* Improved generalization and reduced error by fine-tuning depth and number of estimators

**11. Predictions on New Data**

Tested with sample input:

* Example: Age=45, BMI=32, Smoker=Yes → **Predicted Charges ~ ₹25,000+**
* Example: Age=30, BMI=22, Smoker=No → **Predicted Charges ~ ₹4,000 - 6,000**

**12. Deployment Plan**

1. Save trained model using joblib
2. Build Flask/Streamlit API for prediction
3. Host on cloud (Heroku / AWS / Azure) for real-time use

**13. Conclusions**

1. **Smoker, BMI, and Age** are the strongest predictors of medical charges.
2. **Tree-based models (Random Forest, Gradient Boosting)** outperform Linear Regression.
3. **Gradient Boosting** is the best choice for deployment.
4. Insurance firms can use this model to **design premium pricing** strategies and forecast claims.

**14. Future Enhancements**

* Handle **outliers** with robust regression
* Collect **more real-world data** to improve generalization
* Add new features: pre-existing conditions, exercise habits, etc.
* Deploy interactive web app for real-time predictions