Data Analysis and Prediction on Insurance Dataset

1. Introduction

- **Objective:** Analyze and predict insurance charges based on various factors.
- **Dataset:** insurance.csv with 1338 entries.
- **Features:** age, sex, bmi, children, smoker, region, charges.
- **Target Variable:** charges (insurance cost).
- **Analysis Goals:** Identify key factors and build predictive models.
- **Tools Used:** pandas, matplotlib, scikit-learn.

2. Loading and Exploring the Dataset

- Loaded dataset using pandas.
- Displayed first few rows for initial exploration.
- Identified features and their data types.
- Checked for missing values.

3. Data Preprocessing

- Converted categorical values to numerical using LabelEncoder.
- Transformed sex, smoker, and region.
- Ensured data consistency post-transformation.
- Displayed transformed data.

4. Correlation Analysis

- Visualized correlations using a heatmap.
- Strong correlation: smoker and charges.
- Moderate correlations: age and bmi with charges.
- Weaker correlations: children, sex, region with charges.
- Interpreted key findings.

5. Data Visualization

- Created a pair plot to visualize feature relationships.
- Calculated and interpreted skewness and kurtosis for each feature.
- Noted high skewness and kurtosis for charges.

6. Data Preparation

- Selected features: age, sex, bmi, children, smoker, region.
- Target variable: charges.
- Split data into training (80%) and testing (20%) sets.
- Ensured reproducibility with random_state=42.

7. Model Training and Prediction

- **Linear Regression:** Trained and evaluated.
- **Support Vector Regression (SVR):** Trained and evaluated.
- **Ridge Regression:** Trained and evaluated.
- **Random Forest Regressor:** Trained and evaluated.
- Compared performance metrics: MSE, R2 Score.

8. Hyperparameter Tuning

- Used GridSearchCV for SVR and Random Forest Regressor.
- Tuned parameters: {'C': [0.1, 1, 10], 'gamma': [1, 0.1, 0.01]} for SVR.
- Tuned parameters: {'n_estimators': [10, 50, 100], 'max_features': ['auto', 'sqrt']} for Random Forest.
- Selected best parameters and noted performance improvements.

9. Model Performance Comparison

- Compared Mean Squared Error (MSE) and R2 Score for all models.
- Visualized performance using bar charts.
- Highlighted Random Forest Regressor as the best-performing model after tuning.

10. Conclusion

- **Key Findings:** Smoker status significantly impacts insurance charges. Random Forest Regressor performed best after tuning.
- **Model Insights:** Varying performance of Linear Regression, SVR, Ridge Regression.
- **Future Work:** Explore additional features and advanced models. Implement cross-validation for robust evaluation.
- **Applications:** Potential use in the insurance industry for pricing policies.
- **Closing Remarks:** Emphasized the importance of data analysis and predictive modeling in decision-making.