Medical Costs Dataset Project Report

Introduction

This project analyzes the Medical Insurance Cost Dataset to explore the relationships between medical insurance charges and demographic and health-related variables, including age, sex, BMI, smoking habits, number of children, and region. The objective is to understand how these factors influence medical costs and to build predictive models to estimate charges. The dataset, sourced from "insurance.csv," contains 1338 records and 7 columns, with "charges" as the continuous target variable. The analysis involves data cleaning, visualization, transformation, and predictive modeling using Linear Regression and Random Forest Regressor.

Methodology

- ### Data Loading and Initial Exploration

  Dataset: Loaded using pandas from "insurance.csv" with 1338 records and 7 columns: age (int64), sex (object), bmi (float64), children (int64), smoker (object), region (object), and charges (float64).

  Initial Checks:

- ✓ No missing values were found (confirmed via `med.isnull().sum()`).
- ✓ One duplicate record was identified and removed, reducing the dataset to 1337 records (`med.drop_duplicates()`).
- ✓ Data types: Numerical (age, bmi, children, charges) and categorical (sex, smoker, region).
- ✓ Summary statistics (`med.describe()`) revealed:
- ✓ Age: Mean ~39.21 years, range 18–64 years.
- ✓ BMI: Mean ~30.54, range 15.96–46.75.
- ✓ Children: Mean ~1.10, range 0–5.
- ✓ Charges: Mean ~9.10 (log-transformed), range 7.02–11.04 (log scale), right-skewed before transformation.

- ### Data Preprocessing

- ✓ Handling Duplicates: Removed one duplicate record to ensure data integrity.
- ✓ Categorical Encoding: Converted categorical variables (sex, smoker, region) to numerical values using Label Encoding.

- ✓ **Transformation:**
  - Applied log transformation to the "charges" column to address right-skewness.
  - Removed BMI outliers using IQR-based filtering to improve data quality.

- ✓ Feature Scaling: Standardized numerical features (age, bmi, children, sex, smoker, region) using `StandardScaler` to ensure uniformity for modeling.

- **Exploratory Data Analysis (EDA)**

✓ **Univariate Analysis:**

- **Categorical Variables:**
  - Sex: Nearly balanced distribution (675 males, 662 females).
  - Smoker: Most individuals are non-smokers (79.6% non-smokers vs. 20.4% smokers).
  - Region: Southeast region has the highest representation.
  - Visualized using pie and bar plots.

- **Numerical Variables:**
  - Analyzed distributions of age, bmi , and charges using histograms and box plots.
  - Charges were right-skewed, justifying log transformation.
  - BMI outliers were identified and handled.

✓ **Bivariate Analysis:**
- Smoker vs. Charges: Smokers incur significantly higher charges, indicating a strong influence.
- Age vs. Charges: Positive correlation; older individuals have higher charges.
- BMI vs. Charges: Higher BMI, especially among smokers, correlates with increased charges.
- Region and Sex: Minimal impact on charges compared to smoking status, age, and BMI.
- Visualizations included scatter plots (age vs. charges, colored by intensity) and BMI analysis with smoking status.

Modeling

- **Data Splitting:** Split data into training (80%, 1062 records) and testing (20%, 266 records) sets using `train_test_split` with `random_state=40`.

- **Models Used:**

✓ **Linear Regression:**
- Trained on standardized features to predict log-transformed charges.
- Evaluation metrics:
  - R^2: 80.51%
  - Mean Absolute Error (MAE): 0.2725
  - Mean Squared Error (MSE): 0.1682
- Coefficients indicate smoker status (0.6145) and age (0.4747) as strong predictors.

✓ **Random Forest Regressor:**
- Trained with 50 estimators (`n_estimators=50`).
- Evaluation metric:
  - R^2: 84.95%
- Outperformed Linear Regression, with slight variations (1–2% decrease) when using fewer estimators.

Results

- **Data Characteristics:**

✓ The dataset is clean with no missing values and minimal duplicates.
✓ Charges are heavily influenced by smoking status, age, and BMI, with smokers paying significantly more.
✓ Region and sex have minimal impact on charges, while the number of children shows no strong correlation.

- **Model Performance:**

✓ Linear Regression: Achieved an R^2 of 80.51%, indicating a good fit but limited by linear assumptions.
✓ Random Forest Regressor: Achieved a higher R^2 of 84.95%, demonstrating better predictive accuracy, likely due to its ability to capture non-linear relationships.
✓ Random Forest performance was robust, with minimal sensitivity to the number of estimators (50 vs. 100).

Discussion

- **Key Insights:**
  - ✓ Smoking Status: The strongest predictor of medical costs, with smokers incurring drastically higher charges due to health risks.
  - ✓ Age: Older individuals face higher charges, likely due to increased healthcare needs.
  - ✓ BMI: Higher BMI, particularly among smokers, is associated with elevated costs, reflecting obesity-related health issues.
  - ✓ Children, Region, and Sex: These factors have limited influence on charges, suggesting they are less critical for cost prediction.

- **Model Comparison:**
  - ✓ Random Forest outperformed Linear Regression, capturing complex interactions between variables.
  - ✓ The higher R^2 (84.95%) of Random Forest indicates better generalization to unseen data.
  - ✓ Linear Regression's coefficients provide interpretability, highlighting smoker status and age as key drivers.
- **Data Quality:** Preprocessing steps (duplicate removal, outlier handling, log transformation, and standardization) improved model performance and analysis r2_score.

Conclusion

The analysis successfully identified smoking status, age, and BMI as the primary drivers of medical insurance costs, with Random Forest Regressor providing superior predictive accuracy (R^2 = 84.95%) compared to Linear Regression (R^2 = 80.51%). The dataset was well-prepared through cleaning and transformation, ensuring robust insights. Future work could explore additional models (e.g., Gradient Boosting) or feature engineering (e.g., interaction terms) to further improve predictions. These findings can inform insurance pricing strategies and health policy decisions, emphasizing smoking cessation and BMI management to reduce medical costs.

References

- Dataset: "insurance.csv" (Kaggle).
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn.