Miraj Rahman Sanjid

CS 505 Data Mining

Octavio Juarez Espinosa

Date: 3/4/2025

**Medical Insurance Cost Prediction Report**

**Introduction**  
This report focuses on predicting medical insurance costs using linear regression techniques. The dataset includes demographic and health-related variables such as age, BMI, smoking status, and region. The goal is to predict the medical charges based on these features.

**Data Preprocessing**

* The dataset was inspected for missing values, and rows with null values were dropped.
* Categorical variables (sex, smoker, region) were encoded using one-hot encoding to convert them into numerical form.
* Numerical features (age, bmi, children) were standardized using StandardScaler to ensure uniform scaling across the data.
* The dataset was split into training (80%) and testing (20%) sets for model evaluation.

**Model Training**  
Several linear regression models were implemented:

* Standard Linear Regression
* Ridge Regression (regularization to prevent overfitting)
* Lasso Regression (feature selection through coefficient shrinkage)
* Elastic Net (a combination of Lasso and Ridge)

The models were trained on the preprocessed training set, and their performance was evaluated on the testing set.

**Model Evaluation**  
The models were assessed using the following evaluation metrics:

* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* R-squared (R²)

A scatter plot was created to compare the predicted values against the actual values for the Linear Regression model.

**Interpretation and Analysis**

1. Feature Influence
   * Smoker status had the most significant impact on medical costs, with smokers showing a strong positive effect on charges.
   * Other important features included age and BMI, which also contributed notably to the prediction.
2. Model Performance
   * The R² value indicates the proportion of variance explained by the model. Higher R² values suggest better model fit.
   * MAE and RMSE measure prediction error, with lower values indicating better performance.
   * Ridge and Elastic Net provided better performance due to regularization, while Lasso helped with feature selection.
3. Comparison of Regression Models
   * Ridge Regression helped prevent overfitting by regularizing the coefficients but retained all features.
   * Lasso Regression zeroed out less important features, aiding interpretation by selecting only relevant variables.
   * Elastic Net combined the strengths of both Ridge and Lasso, providing a balance between feature selection and regularization.

**Conclusion**

* The linear regression models were effective at predicting medical insurance costs, with smoking status being the most influential factor.
* Regularization methods (Ridge, Lasso, and Elastic Net) improved model performance and interpretability.
* Future work could involve exploring more advanced models such as decision trees or neural networks to improve prediction accuracy.