Group 45: Health Insurance Cost Prediction

Name=Bikal Bista(A43323)

1.Objective

- **Goal**: Develop a machine learning model using Support Vector Machines (SVM) to predict healthcare costs for new customers.
- **Outcome**: Provide accurate and fair insurance premium estimations.

2.Dataset Overview

- **Training Data**: 2,215 instances with features like gender, marital status, area of residence, BMI, smoking status, age class, and healthcare costs.
- **Test Data**: 550 instances with only features (costs unknown).

3.Methodology

1. **Data Preprocessing**:
    - Categorical features encoded using LabelEncoder.
    - Data scaled using StandardScaler.
    - Dataset split: 80% training and 20% testing.
2. **Model Training**:
    - Algorithm: Support Vector Regressor (SVR)
    - Key Parameters: Kernel = 'rbf', C = 100, Gamma = 0.1, Epsilon = 0.1
3. **Feature Importance Analysis**:
    - Technique: Permutation Importance
    - Tool: Scikit-learn’s permutation_importance function

4.Model Evaluation

- **Performance Metrics**:
    - R^2 Score: 0.19

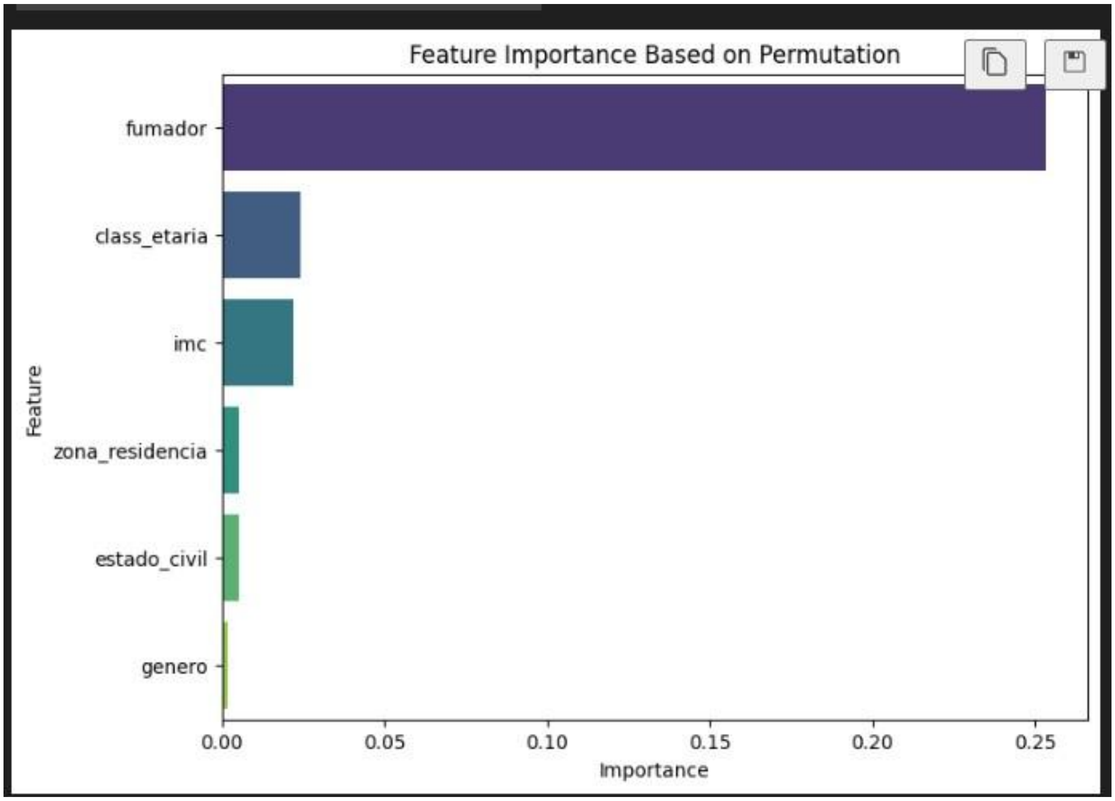
- Mean Squared Error (MSE): 89,513,449.68

5.Observations:

- Model shows moderate predictive performance.
- Smoking status (fumador) is the most impactful feature.

6.Feature Importance

- **Top Influential Features**:
  1. Smoking Status (fumador)
  2. Age Class (class_etaria)
  3. Body Mass Index (imc)
- **Visualization**: The bar chart below illustrates the importance of each feature:


Feature Importance Based on Permutation
fumador
class_etaria
imc
zona_residencia
estado_civil
genero
Feature
Importance