Machine Learning Course

Assignment 1

Medical Cost Dataset

Description

The Medical Cost Dataset contains demographic and health-related features of patients, such as age, BMI, smoking habit, number of children, region, etc., along with their corresponding medical costs.

Attributes

- **Age**: Age of the patient (years)
- **Sex**: Gender of the patient (male or female)
- **BMI**: Body mass index ($\frac{\text{weight}}{\text{height}^2}$)
- **Children**: Number of children/dependents covered by health insurance
- **Smoker**: Smoking habit of the patient (yes or no)
- **Region**: Geographic region of the patient (northeast, southeast, southwest, northwest)
- **Charges**: Medical costs incurred by the patient

Objective

The objective of this dataset is to perform a regression task to predict the medical costs incurred by patients based on their demographic and health-related features.

Source

You can download this dataset from here.

Tasks

1. **Data preprocessing**: Encode categorical variables, handle missing values, normalize/standardize features if necessary.
2. **Exploratory Data Analysis (EDA)**: Analyze the distribution of features, correlations, etc.

3. Feature selection/engineering: Select relevant features and/or create new features if needed.
4. **Model selection**: Experiment with various regression algorithms (e.g., Linear Regression, Random Forest Regression, Gradient Boosting, etc.).
5. **Model evaluation**: Evaluate the models using appropriate evaluation metrics (e.g., Mean Absolute Error, Mean Squared Error, R-squared).
6. **Hyperparameter tuning**: Fine-tune the parameters of the best performing model.
7. **Final model selection**: Select the best model based on evaluation results.
8. **Model interpretation**: Interpret the results and understand the factors influencing medical costs.

Diabetes Dataset

Description

The Diabetes Dataset contains various health-related features of patients, such as glucose levels, insulin levels, BMI, age, etc., and a binary target variable indicating whether the patient has diabetes or not.

Attributes

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration after 2 hours in an oral glucose tolerance test
- **BloodPressure**: Diastolic blood pressure (mm Hg)
- **SkinThickness**: Triceps skin fold thickness (mm)
- **Insulin**: 2-Hour serum insulin (mu U/ml)
- **BMI**: Body mass index ($\frac{\text{weight}}{\text{height}^2}$)
- **DiabetesPedigreeFunction**: Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
- **Age**: Age of the patient (years)
- **Outcome**: Target variable (0 if non-diabetic, 1 if diabetic)

Objective

The objective of this dataset is to perform a classification task to predict whether a patient has diabetes or not based on the given health-related features.

Source

You can download this dataset from here.

Tasks

1. **Data preprocessing**: Handle missing values, normalize/standardize features if necessary.
2. **Exploratory Data Analysis (EDA)**: Analyze the distribution of features, correlations, etc.
3. **Feature selection/engineering**: Select relevant features and/or create new features if needed.
4. **Model selection**: Experiment with various classification algorithms (e.g., Logistic Regression, Random Forest, SVM, etc.).
5. **Model evaluation**: Evaluate the models using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score).
6. **Hyperparameter tuning**: Fine-tune the parameters of the best performing model.
7. **Final model selection**: Select the best model based on evaluation results.
8. **Model interpretation**: Interpret the results and understand the importance of features in predicting diabetes.