MACHINE LEARNING DOCUMENTATION

Table of Contents

Contents

First: Numeric Dataset

1. Dataset information:

We focused on **Healthcare Insurance Expenses** from Kaggle and its link: https://www.kaggle.com/datasets/arunjangir245/healthcare-insurance-expenses/data

Its idea is: **Develop predictive models for estimating healthcare expenses.**

The columns (features) for input are the following:

1. **Age**: The insured person's age.
2. **Sex**: Gender (male or female) of the insured.
3. **BMI (Body Mass Index)**: A measure of body fat based on height and weight.
4. **Children**: The number of dependents covered.
5. **Smoker**: Whether the insured is a smoker (yes or no).
6. **Region**: The geographic area of coverage.
7. **Charges**: The medical insurance costs incurred by the insured person.

It consists of 1338 rows and 7 columns shown previously.

The input train consists of 935 rows and 3 columns while the input test consists of 402 rows and 3 columns.

2. Implementation Details:

2.1. **Feature Extraction**:

At feature extraction, we extracted 3 features (columns) from 6 columns which are: **smoker, age & BMI**. The dimensions of resulted features are 1337 row × 3 columns.

3. Results details:

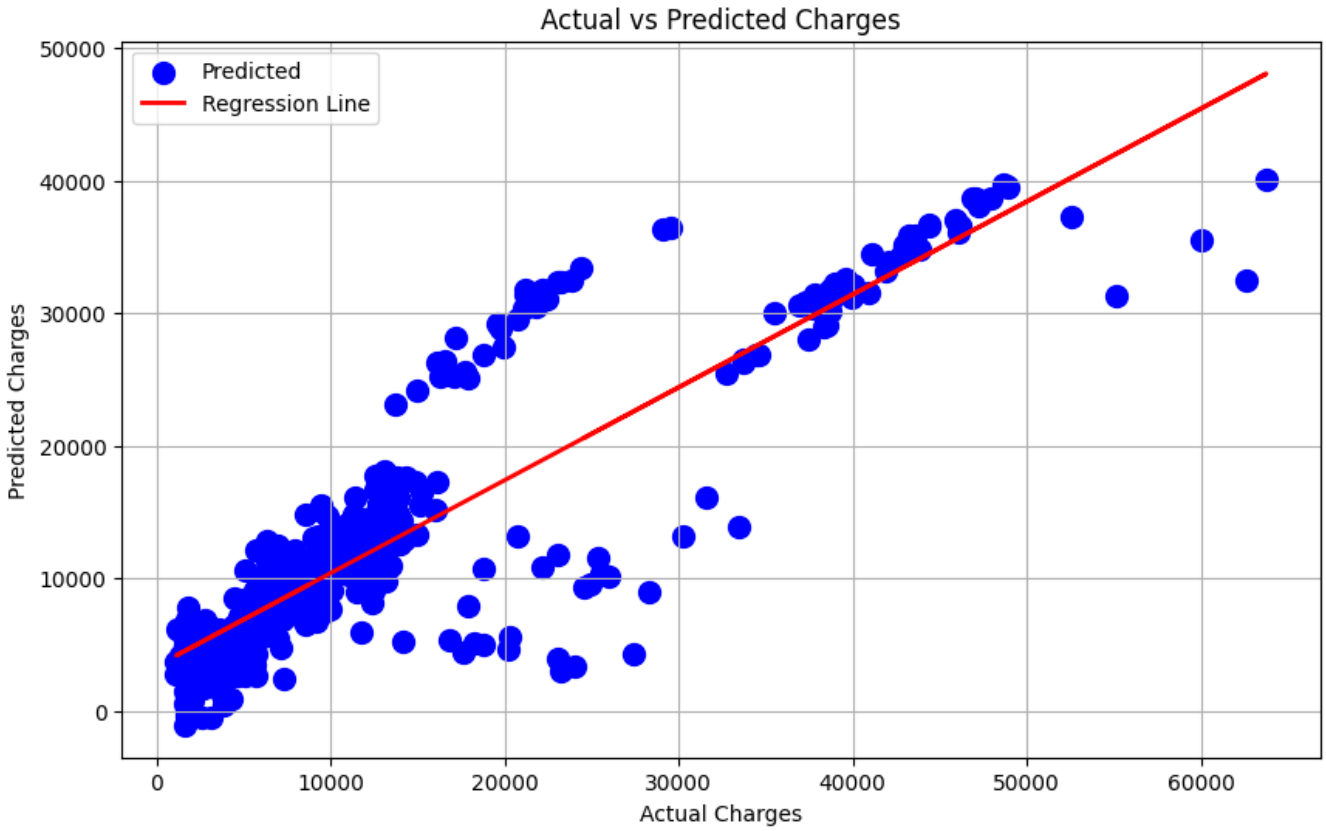
3.1. Linear Regression:

3.1.1. **R2 Score:** 0.770150306428094 3

3.1.2. **Mean Absolute Error:** 4219.085290485933

3.1.3. **Mean Squared Error:** 39331352.77834985

3.1.4. Plot Actual charges bs Predicted Charges



3.2. KNN:

3.2.1. R2 Score: 0.22153314883467312

3.2.2. Mean Absolute Error: 7352.568930097014

3.2.3. Mean Squared Error: 133209463.42639403

3.3. Comparison between Linear Regression and KNN according to bar plot