Academic Year	Module	Assessment Number	Assessment Type
2025	Concepts and Technologies of AI	3	Report Writing

An End- to- End Machine Learning Project on Regression and Classification Task

Student ID: 2408646

Tutor Name: Siman Giri

Sudent Name : Prerit Pandey

Section: L4CG1

Submitted on: 11/02/2025

Regression Analysis Report

Abstract

Purpose The aim of this report is to create a regression model from several demographic and health-related variables that would best predict the insurance charges of a continuous variable. Some of the applied regression techniques used in enhancing predictive accuracy involve feature selection, model building, and tuning of hyperparameters.

Approach

EDA, development of the regression model, tuning of hyperparameters, feature selection, and others are discussed for some demographic and health-related attributes in this dataset. Later, two regression models have been developed using Linear Regression and Ridge Regression, and these models were then evaluated against each other based on performance metrics R^2 and RMSE.

Key Results

The Linear Regression model had the best performance, with an R^2 score of 0.78 and a Root Mean Squared Error of 5,796. Based on the feature importance analysis, three of the most important features-age, BMI, and smoker status-feature in insurance charge prediction.

Conclusion

These results provide further verification that the developed model yields a dependable forecast of the insurance charge. Major predictors identified are age, BMI, and smoking status. Further refinement could be achieved with more sophisticated algorithms such as XGBoost.

1. Introduction

1.1 Problem Statement

The cost of insurance varies from person to person and depends on many factors, demographic or health-wise. This study will develop a predictive model that estimates the insurance charges of an individual based on his or her characteristics, such as age, BMI, smoker or not, among other features.

1.2 Dataset Description

The dataset consists of 1,338 entries along with 7 feature:

• Age (Continuous)
• BMI (Continuous)
• Children (Discrete, number of dependents)
• Smoker (Categorical (yes/no))
• Sex (Categorical: Male or Female)
- Region (Categorical: Geographical area)

- Charges

1.3 Objective

The goal is to make a robust regression model which is able to predict insurance charges as close as possible and also identify the most influencing factors.

2. Methodology

2.1 Preprocessing

Missing Values, Encoding Categorical Variables and Feature Scaling had been applied just to keep all continuous variables in the same scale.

2.2 EDA

EDA techniques used included: correlation analysis, scatter plots, distribution analysis, and boxplots to show outliers and trends. 2.3 Model Building Linear Regression Ridge Regression o Dataset splitting into an 80-20 ratio for training-testing purposes was considered. 2.4 Model Evaluation Linear Regression: R^2 Score: 0.78, RMSE: 5,796 Ridge Regression: R^2 Score: 0.78, RMSE: 5,796 2.5 Hyperparameter Optimization

GridSearchCV was applied at Ridge Regression and chose alpha=10.0 - That was not a great outperformance

2.6 Feature Selection

While developing feature selection, following features will explain insurance charges better : Age, BMI, smoker status and no of children.

3. Conclusion

3.1 Key Findings

Status of smoking is the variable most influential in causing change in insurance charge. Next important variables were BMI and age. Linear regression gave the best performance out of all regression models with R^2 score as 0.78.

3.2 Challenges

Outliers, feature interaction and small dataset attributes took a toll on model performance

3.3 Future Work

Further improvement will be done on the model's accuracy with higher-order models such as XGBoost, feature engineering, and outlier handling.