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 Task

Student ID: 2408646

Tutor Name: Siman Giri

Sudent Name : Prerit Pandey

Section: L5CG1

Submitted on: 12/02/2025

Regression Analysis Report

Abstract

Purpose

This report aims to develop a regression model to predict a continuous variable—insurance charges—based on demographic and health-related factors. Regression techniques, including feature selection, model building, and hyperparameter tuning, were employed to improve predictive accuracy.

Approach

The dataset contains demographic and health-related attributes. The methodology includes Exploratory Data Analysis (EDA), regression model development, hyperparameter tuning, and feature selection. Two regression models—Linear Regression and Ridge Regression—were implemented, and their performance was evaluated using R-squared (R^2) and Root Mean Squared Error (RMSE) metrics.

Key Results

The best-performing model, Linear Regression, achieved an R^2 score of 0.78 and an RMSE of 5,796. Feature importance analysis identified age, BMI, and smoking status as the most influential factors in predicting insurance charges.

Conclusion

The results indicate that the developed model provides a reliable prediction of insurance charges, with significant predictors being age, BMI, and smoking status. Future enhancements may involve incorporating more complex models like XGBoost to improve performance.

1. Introduction

1.1 Problem Statement

The cost of insurance varies widely among individuals due to multiple demographic and health-related factors. This study aims to build a predictive model that estimates insurance charges based on a person's characteristics, such as age, BMI, smoking status, and other features.

1.2 Dataset

The dataset comprises 1,338 records with seven features:
- Age (Continuous)

- BMI (Continuous)
- Children (Discrete, number of dependents)
- Smoker (Categorical: Yes/No)
- Sex (Categorical: Male/Female)
- Region (Categorical: Geographical region)
- Charges (Target variable, Continuous)

1.3 Objective

The primary objective is to develop a robust regression model to predict insurance charges as accurately as possible while identifying the most influential factors.

2. Methodology

2.1 Data Preprocessing

Handling Missing Values, Encoding Categorical Variables, and Feature Scaling were applied to ensure uniformity among continuous variables.

2.2 Exploratory Data Analysis (EDA)

EDA techniques included Correlation Analysis, Scatter Plots, Distribution Analysis, and Boxplots to detect outliers and trends.

2.3 Model Building

Two regression models were implemented:
- Linear Regression
- Ridge Regression

Dataset was split into an 80-20 ratio for training and testing.

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 to Ridge Regression to fine-tune the regularization parameter (alpha), which was found to be 10.0, though it did not improve performance significantly.

2.6 Feature Selection

The most significant features influencing insurance charges were Age, BMI, Smoking Status, and Number of Children.

3. Conclusion

3.1 Key Findings

Smoking status has the most significant impact on insurance charges, followed by BMI and age. Linear Regression performed well, achieving an R^2 score of 0.78.

3.2 Challenges

Outliers, feature interactions, and limited dataset attributes impacted model performance.

3.3 Future Work

Advanced models like XGBoost, feature engineering, and handling outliers can further enhance model accuracy.