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

An End-to-End Machine Learning Project on Classification Task.

Student ID: 2408646

Tutor Name: Siman Giri

Sudent Name : Prerit Pandey

Section: L5CG1

Submitted on: 12/02/2025

Classification Analysis Report

Abstract

This report describes the process that was followed to predict a categorical variable.

Methodology: Since the dataset consisted of demographic and health-related data, EDA, modeling through Random Forest, hyper-parameter tuning, and feature selection were done.

Key findings: The accuracy provided by the Random Forest model is 94.74%.

Conclusion: The model performs really great regarding precision. Most impacting features of the prediction include the status of smoking, and the BMI of the person.

1. Introduction

1.1 Problem Statement

The goal is to classify people based on health-related features. Example applications can be medical diagnosis or spam detection. 1.2 Description of Dataset This dataset consists of 1,338 rows and 7 columns of features including age, sex, BMI, children, smoker or non-smoker, region, insurance charges. 1.3 Goal Develop a predictive model to classify a person based on the given features. 2. Methodology 2.1 Preprocessing Cleaning for missing values and outliers. Encoding of categorical variables and normalization of numerical values. 2.2 Exploratory Data Analysis

Examine important relations by the use of a histogram and a correlation matrix. A smoker's status and BMI became major influencing features.

2.3 Model Building

Random Forest classifier with 80-20 train-test split.

2.4 Model Evaluation

Accuracy: 94.74%

Precision, Recall, F1-score: All very high across all classes

2.5 Hyperparameter Optimization

Using GridSearchCV to optimize max_depth = 10, n_estimators = 100

2.6 Feature Selection

Most informative features: Smoking, BMI, worst concavity

3. Conclusion

3.1 Key Findings

The best classification performance of 94.74% has been obtained with Random Forest, where smoking and BMI are the dominant factors in classification.

3.2 Challenges

Due to an imbalance of classes, recall suffered.

3.3 Future Work

Use Gradient Boosting for better performance.