



5CS037 - Concepts and Technologies of AI

Proposal for Final Portfolio Project

Name: Pratisha Bista

Group: L5CG1

University ID: 2408284

Module Leader: Mr. Siman Giri

Tutor: Mr. Siman Giri

Objective of the Proposal

The aim of this proposal is to outline my plan for the final portfolio assignment, utilizing a novel dataset from valid sources. The dataset that I intend to use for classification and regression task aligns with **Sustainable Development Goal 3: Good Health and Well-being** by focusing on **heart disease** prediction and analysis.

Research Questions

Regression Task: How can we predict resting blood pressure (“trestbps”) based on demographic and clinical features like age, cholesterol levels, chest pain type and more?

Classification Task: How accurately can machine learning models classify the presence of heart disease (target variable) based on patient attributes such as age, gender, exercise-induced angina, and other clinical indicators?

Dataset Description

Source: The dataset is fetched from Kaggle datasets and contains clinical and demographic data on cardiovascular health. It is not preloaded in scikit-learn, ensuring compliance with dataset selection requirements. The dataset consists of 1,025 rows and 14 columns.

Key Features:

Independent Variables: Age, sex, chest pain (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), exercise-induced angina (exang), oldpeak, slope, number of major vessels (ca), and thalassemia (thal).

Target Variables: For classification, the target variable is “target” (presence of heart disease: 0 = no, 1 = yes). For regression, the target variable is “trestbps” (resting blood pressure).

Preprocessing Challenges:

Handling missing values, undefined values, invalid zeros, misentered data, encoding categorical variables, and more.

Connection to Sustainable Development Goal (SDG)

Cardiovascular health is a major modern-day health concern. This dataset provides valuable insights into various indicators contributing to heart health, such as cholesterol levels, blood pressure, chest pain types, and exercise-induced factors. Through data analysis, we can better understand their relationships and influence on heart disease risk. The classification model can help identify people at risk of heart disease. The regression model can, in turn, help identify critical indicators such as resting blood pressure.

This aligns directly with SDG 3: Good Health and Well-being, by not only addressing treatment but also promoting preventive healthcare and individual awareness.