

Data Analytics Lab (CS689)

Lab 5

The `health_risk_data` comprises medical records of 800 individuals, each described by a range of biometric and biochemical indicators commonly used to assess the risk of metabolic syndrome, a condition associated with obesity, hypertension, and insulin resistance.

Your objective is to perform a comprehensive multivariate analysis to explore relationships and redundancies among features using visual tools such as pair plots, heatmaps, and facet grids. Based on your findings, determine the redundant features and decide whether applying Principal Component Analysis (PCA) would be beneficial for reducing feature redundancy. If PCA is applied, use a Scree Plot to identify the optimal number of principal components to retain, and then evaluate how PCA influences the classification accuracy and computation time of a Random Forest classifier.

Dataset Description:

Category	Feature Name	Description
Demographic	<code>age</code>	Age in years
Body Metrics	<code>bmi</code>	Body Mass Index (kg/m^2)
	<code>waist_circumference</code>	Waist size (cm) — another obesity indicator
Cardiovascular	<code>systolic_bp</code>	Systolic blood pressure (mmHg)
	<code>diastolic_bp</code>	Diastolic blood pressure (mmHg)
	<code>mean_arterial_pressure</code>	Derived measure combining systolic and diastolic
Metabolic	<code>fasting_glucose</code>	Fasting blood sugar (mg/dL)
	<code>hb1c</code>	Glycated haemoglobin (%)
	<code>glucose_monitor_avg</code>	Average glucose from wearable sensor
Lipid Profile	<code>total_cholesterol</code>	Total cholesterol (mg/dL)
	<code>ldl_cholesterol</code>	“Bad” cholesterol
	<code>hdl_cholesterol</code>	“Good” cholesterol
Heart Rate	<code>resting_hr</code>	Resting heart rate (bpm)
	<code>activity_hr</code>	Average heart rate during light activity
Other variables	<code>device_artifact_signal</code>	Spurious sensor readings from a wearable device
	<code>measurement_variability_index</code>	Unexplained variability across repeated tests

Category	Feature Name	Description
Target	risk_label	Binary outcome: 0 = low risk, 1 = high risk

Objectives:

- a. Perform a multivariate analysis of the given dataset using the visualization techniques discussed in Labs 3 and 4 (e.g., pair plots, heatmaps, and facet grids) to explore relationships among variables.
- b. Identify redundancy among the features by analysing correlations and overlapping patterns observed in the visualizations.
- c. Justify the need for applying PCA by explaining why dimensionality reduction may be preferable to using all features directly in the classification model.
- d. Generate a Scree Plot and determine the number of principal components (PCs) required to capture at least 95% of the total variance in the dataset.
- e. Train and evaluate a Random Forest classifier on both the original and PCA-transformed datasets and compare the results to assess whether PCA leads to any improvement in classification accuracy and computation time.