

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	age	Age in years
Body Metrics	bmi	Body Mass Index (kg/m ²)
	waist_circumference	Waist size (cm) — another obesity indicator
Cardiovascular	systolic_bp	Systolic blood pressure (mmHg)
	diastolic_bp	Diastolic blood pressure (mmHg)
	mean_arterial_pressure	Derived measure combining systolic and diastolic
Metabolic	fasting_glucose	Fasting blood sugar (mg/dL)
	hba1c	Glycated haemoglobin (%)
	glucose_monitor_avg	Average glucose from wearable sensor
Lipid Profile	total_cholesterol	Total cholesterol (mg/dL)
	ldl_cholesterol	“Bad” cholesterol
	hdl_cholesterol	“Good” cholesterol
Heart Rate	resting_hr	Resting heart rate (bpm)
	activity_hr	Average heart rate during light activity
Other variables	device_artifact_signal	Spurious sensor readings from a wearable device
	measurement_variability_index	Unexplained variability across repeated tests

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

Objectives:

- 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.
- Identify redundancy among the features by analysing correlations and overlapping patterns observed in the visualizations.
- Justify the need for applying PCA by explaining why dimensionality reduction may be preferable to using all features directly in the classification model.
- 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.
- 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.