Populaire et Démocratique Algérienne 6
République
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

---

Ecole Supérieure d'Informatique - Sidi Bel Abbes
المدرسة الوطنية العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس
2éme Année - Second Cycle

PCA

Data Analysis

Members:

- Benahmed Firdaws
- Benghenima Hafsa
- Ghandouz Amina
- Meflah Yousra

Dr. Keskes Nabil

Table de matière

• Introduction To PCA:

PCA (Principal Component Analysis) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while retaining as much variance/information as possible. It works by identifying the directions (principal components) in which the data varies the most and projecting the data onto these components. We're using PCA to create new, uncorrelated features (principal components) that are linear combinations of the original features (feature extraction) so we will apply it to the diabetes dataset to reduce its dimensionality and analyze the relationships between features.

• About the dataset

The dataset provided contains health-related measurements for individuals, with the goal of predicting whether a person has diabetes. It includes 8 features and 1 target variable (outcome). Each sample represents a patient, and the columns represent various health metrics. We are using the dataset for a binary classification task, where the goal is to predict the presence or absence of diabetes based on the given features.

- **About features**

-Glucose: measures plasma glucose concentration after a 2-hour oral glucose tolerance test (148 mg/cl), high levels may indicate diabetes or prediabetes.

-**BloodPressure**: represents diastolic blood pressure (72 mm Hg), it indicates cardiovascular health, with abnormal levels potentially signaling hypertension or other conditions.

-**SkinThickness**: measures the thickness of the skin fold at the triceps (35 mm), it estimates body fat percentage, which is linked to diabetes risk.

-**Insulin**: represents 2-hour serum insulin levels (, 0 mu U/mL), it reflects the body's ability to regulate blood sugar, with low levels indicating insulin resistance.

-**BMI**: calculates body mass index (33.6), a measure of body fat based on weight and height, high BMI is associated with obesity and diabetes risk.

-**DiabetesPedigreeFunction**: scores the likelihood of diabetes based on family history (0.627), it indicates genetic predisposition to the disease.

-**Age**: represents the patient's age (50 years), older age is a known risk factor for diabetes.

-**Outcome**: the target variable, indicates whether the patient has diabetes (1) or not (0).

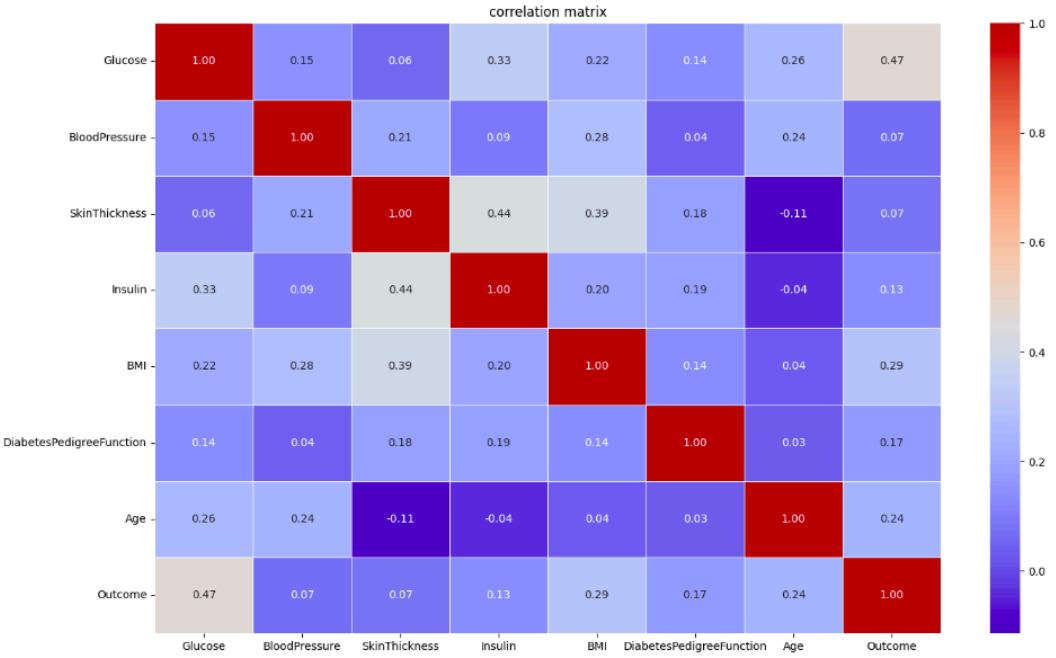
• Proposed solution:

- **Used metrics:**

-**Eigenvalues**: the variance explained by each PC.
-**Percentage of Inertia**: the proportion of variance explained by each PC.
-**Cumulative Percentage**: the cumulative variance explained up to each PC.
-**Contributions**: the influence of original features on PCs.
-**Representational Qualities**: how well individuals/variables are represented by PCs.
-**Correlation Circle**: visual representation of feature correlations with PCs.

- **Correlation matrix:**

it shows the pairwise correlation coefficients between all features in the dataset in order to identify relationships between variables, such as which features are strongly positively or negatively correlated, where it provides ranges from -1 to 1:
  - 1 refers to perfect positive correlation
  - -1 refers to perfect negative correlation
  - 0 means no correlation



- **What does the sum of the eigenvalues correspond to:**

it corresponds to the total variance in the dataset, each eigenvalue represents the variance explained by a principal component, if the eigenvalues are λ_1, λ_2 …λ_n then

Total Variance = $\lambda_1 + \lambda_2 + \ldots + \lambda_n$

- **justify the choice of the first two PC:**

they are chosen because they explain the majority of the variance in the dataset (75% cumulative inertia), this reduces dimensionality while retaining most of the information, the first two eigenvalues are much larger than the rest, which indicates that the first two PC explain most of the variability in the data.

- **Which Subtests Are Most Strongly Correlated:**

identify the features that are most strongly correlated from the correlation matrix,

the strongest positive correlations:
Glucose & Outcome: 0.466581
SkinThickness & Insulin: 0.436783

the strongest negative correlations:

SkinThickness & Age: -0.11397

- **Graphical presentation of variables in PC1 and PC2:**

the variables are well-represented in the PC1, PC2 plane because:

-Variance Captured: PC1 and PC2 explain the majority of the dataset's variance, ensuring meaningful representation.

-Correlation Circle: Variables are plotted as vectors, where longer vectors indicate stronger contributions to PC1 and PC2, and angles between vectors show their relationships (positive or negative correlation).

-Simplified Interpretation: The 2D plane reduces complexity, making it easier to visualize and interpret variable relationships.