Challenge Campus Biomedico

Supervised Clustering with target variable: an application in Teleassistance

26/03/2024

ENGINEERING

Context: Telemedicine

Telemedicine is a medical support service which allows the remote interaction, through an electronic device, between patients and health professionals. PNT (Piattaforma Nazionale di Telemedicina) aims at **governing and monitoring** the telemedicine processes conducted at **regional level**. The main objective is to harmonize these processes within the **national healthcare system**, coordinating them with the **specific digital ecosystem** of each region.
To achieve this goal, resources and services are provided to facilitate processes **integration and development**; being able to expand as much as possible in the national territory.

Some of the Telemedicine goals are:

- Simplifying chronic illnesses management
- Boosting dehospitalization
- Improving clinical quality and the access to healthcare services
- Providing health professionals with new innovative tools

Context: TeleAssistenza

Among the various services offered by the platform, to present the challenge, we will consider the Teleassistance service.
Teleassistance is a service that involves **medical visits between patients and healthcare professionals** (nurses, psychologists, educators), aimed to provide check-ups and treatment's assistance.

The platform tracks every provided assistance and acquires all necessary data to ensure appropriate **historical documentation**.

Challenge: Cluster Associated with an Outcome Variable

The goal of the challenge is to profile patients taking account of their **contribution to Teleassistance service increase**.
It's important to identify **common patterns and behaviours** according the **increase of teleassistances due to standard patients**.

The approach involves **identifying groups of patients** related to a **particular outcome** variable/objective.
In other words, it entails identifying **patient's groups with common patterns** or **similar behaviours** in accordance with the target variable **(y=incremento_teleassistenze)**.
Next, the differences among patients from various group of increment are analyzed, to understand which **features lead** to the teleassistances increase.

To identify those groups (clusters), **advanced clustering methods** are required to consider both patient characteristics and the outcome variable (incremento_teleassistenze).
In particular, **supervised clustering techniques** are requested for the challenge.

Analysis purpose

Through the analysis of this patients profiling, it will be possible to identify and extract **relevant information** regarding patients.
Going beyond simple aggregate statistics, it's possible to obtain a **more detailed and personalized** view of their behaviours.
This is a fundamental process for understanding which **factors that influence growth or change** in the utilization of this remote healthcare service.

Supervised Clustering 1/2

Traditional clustering is typically applied in an **unsupervised** learning framework using particular **error functions**, e.g. an error function that **minimizes/maximizes the distances/likelihood** inside a cluster keeping clusters tight.

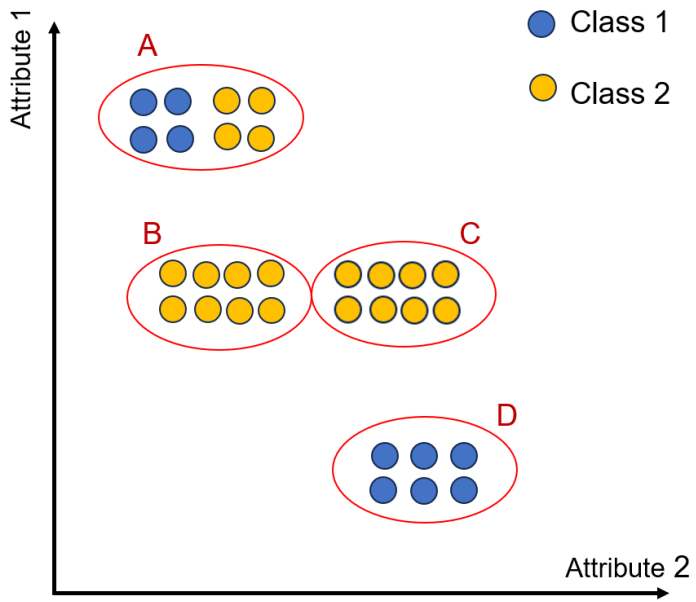
Supervised clustering (SC) uses labeled data as **prior knowledge** or **constraints** to guide the unsupervised clustering of the remaining unlabeled data.

Utilizing the strengths of both supervised and unsupervised approaches, supervised clustering (SC) can enhance the **quality**, **consistency**, and **interpretability** of clusters when dealing with data that has complex or ambiguous structures.

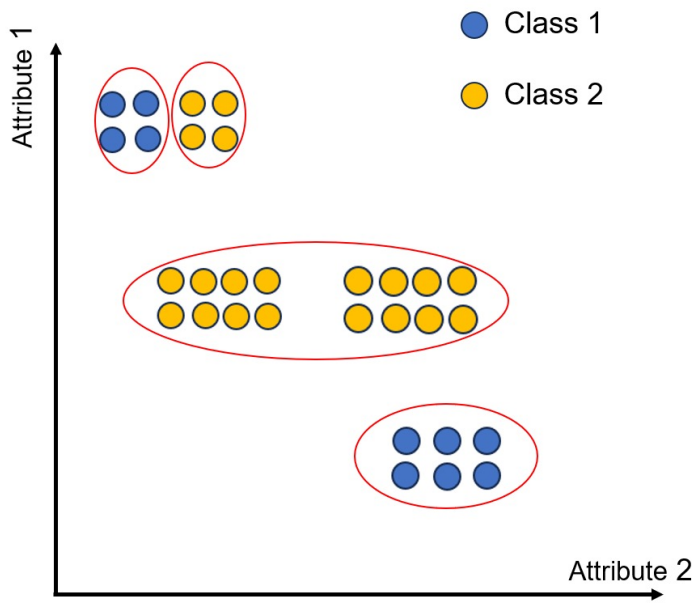
In this challenge, we assume to evaluate the performance of supervised clustering using the following metrics:

- **Class purity:** measured by the percentage of majority class examples in the different clusters of a clustering iteration.
- **Cluster quality:** measured by Silhouette Score using a Euclidean distance function.
- **Number of clusters (k):** preferably to be kept low.

Supervised Clustering 2/2



a. **Traditional Clustering**



Supervised Clustering

Traditional clustering algorithm would identify the four clusters pictured in the upper graph. The reason is that traditional clustering is ignorant with respect the **class membership** of examples. If goal is to generate summaries for classes 1 and 2, the traditional clustering would not be very attractive; since it combined objects from different classes into cluster A, and put examples of the same class in two different clusters B and C.

Supervised clustering algorithm, that maximizes class purity, on the other hand would split traditional cluster A into two clusters. Consequently, traditional clusters B and C would be merged into one cluster, without compromising class purity while reducing the number of clusters.

The cluster prototype, which is an average representation of the points within a cluster, provides useful information to identify characteristics and structure of the cluster itself, and can also be used to identify the target variable importance.

Solution: Algorithmic Flow

Data input

Data Preprocessing

Data Cleaning ← Data Transformation

Data Reduction

Features Extraction

Clustering

Evaluation Metrics

Analysis results

Data input 1/2

The data for this challenge are provided through PARQUET file, which contains the following features:

Nome Variabile	Description	Type
id_prenotazione	Unique identifier of a single Teleassistance	String
id_paziente	Patient's unique identifier code	String
data_nascita	Patient's birth date	String
sesso	Patient's sex	String
regione_residenza	Patient's residence region	String
codice_regione_residenza	Patient's residence region code	String
asl_residenza	Patient's residence ASL	String
codice_asl_residenza	Patient's residence ASL code	String
provincia_residenza	Patient's residence province	String
codice_provincia_residenza	Patient's residence province code	String
comune_residenza	Patient's residence city	String
codice_comune_residenza	Patient's residence city code	String
tipologia_servizio	Typology of offered service from telemedicine platform	String
descrizione_attivita	Description of performed activity	String
codice_descrizione_attivita	Typology of performed activity's	String
data_contatto	Patient's contact date	String

Data input 2/2

Nome Variabile	Description	Type
regione_erogazione	Service's erogation region	String
codice_regione_erogazione	Service's erogation region's code	String
asl_erogazione	Service's erogation ASL	String
codice_asl_erogazione	Service's erogation ASL code	String
provincia_erogazione	Service's erogation province	String
codice_provincia_erogazione	Service's erogation province code	String
struttura_erogazione	Service's erogation facility name	String
codice_struttura_erogazione	Service's erogation facility name's code	String
tipologia_struttura_erogazione	Service's erogation facility typology	String
codice_tipologia_struttura_erogazione	Service's erogation facility typology code	String
id_professionista_sanitario	Healtcare professional erogator's unique identifier code	String
tipologia_professionista_sanitario	Healtcare professional erogator's typology	String
codice_tipologia_professionista_sanitario	Healtcare professional erogator's typology code	String
data_erogazione	Service's erogation date	String
ora_inizio_erogazione	Service's erogation start timestamp (if already permormed)	String
ora_fine_erogazione	Service's erogation end timestamp (if already permormed)	String
data_disdetta	Service's erogation cancellation timestamp (if visit cancelled)	String

Data Preprocessing

Data preprocessing is a set of techniques, fundamental in machine learning, that involves transforming raw data into an **“machine learning understandable" format**. Major tasks in Data Preprocessing are:

- **Data cleaning**
  - Fill in missing data
  - Smooth noisy data
  - Identify or remove outliers
  - Remove duplicates
- **Data transformation**
  - Normalization
  - Aggregation
- **Data reduction**
  - Reduce volume of representations (while producing the same or similar analytical results)
  - Remove redundant columns

Feature Extraction

From existing features, extract additional ones to enhance the data analysis process.

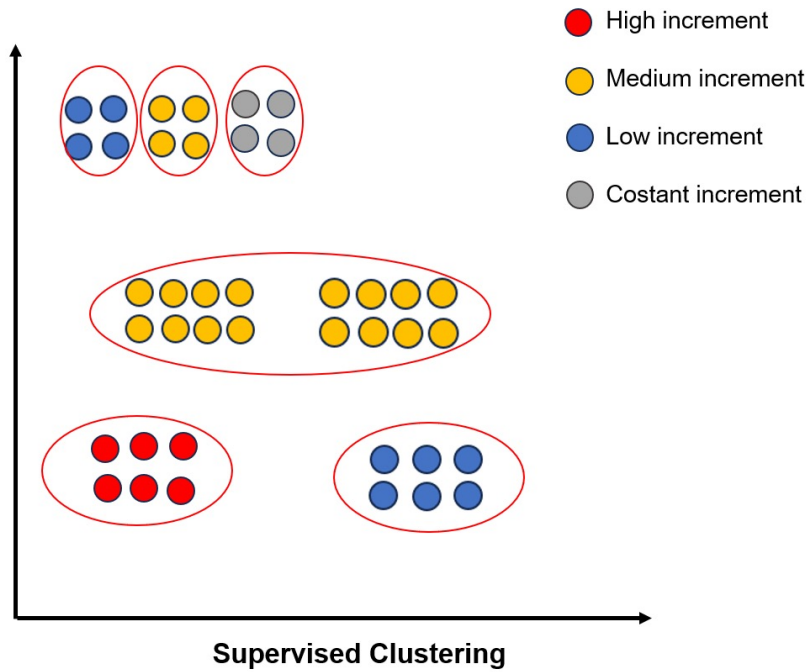
Example:

Variabile Name	Description
eta	Patient's age at the time of the visit
durata_assistenza	Duration of the Teleassistance service provided to the patient
incremento	Difference in the number of Teleassistance services provided, e.g. between corresponding quadrimesters of one year and the next
incremento_teleassistenze	The target variable, calculated from **incremento**

incremento_teleassistenze: target variable must be discretized in 4 classes:

- **Costant** increment
- **Low** increment
- **Medium** increment
- **High** increment

Clustering



Group patients into homogeneous clusters based on their characteristics (common patterns or similar behaviors) and target of interest (incremento_teleassistenze), with the aim of maximizing both the overall **consistency and purity** of the clustering.

Proposed Evaluation Metrics 1/3

Purity: measured the purity of every cluster, which evaluates how much each cluster contains elements of the same class. The metric, for each cluster, identifies the most represented class and counts how many elements belong to it. Next, it normalizes the sum of the obtained values.

Purity is calculated as follows:

$$Purity = \frac{1}{N}\sum_{k} \max_{j} \left|C_k \cap L_j\right|$$

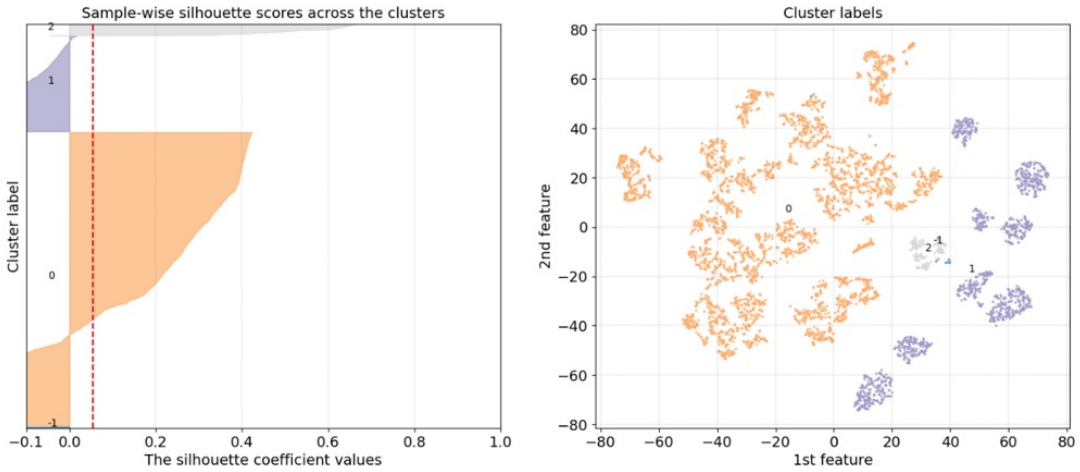
N: total number of elements to be clustered.
k: number of clusters.
C_k: set of elements in cluster k.
L_j : set of elements belonging to class j.
$\left|C_k \cap L_j\right|$: number of elements in cluster k belonging to class j.
$\max_j \left|C_k \cap L_j\right|$: maximum number of elements in C_k belonging to the same class.

Purity varies between 0 and 1; a value of 1 indicates perfect clustering, where each cluster contains only elements from one class.

Proposed Evaluation Metrics 2/3

The Silhouette metric can be used to evaluate the **accuracy** of clustering.

Silhouette Score: a measure of how similar an object is to its own cluster (**cohesion**) compared to other clusters (**separation**).
The silhouette ranges from −1 to +1.



To make it more interpretable, it is often normalized to a range between 0 and 1.
A normalized Silhouette Score of:

- 1 → indicates that the samples are assigned to the correct cluster.
- 0.5 → indicates that the samples are on the boundary between two clusters.
- 0 → indicates that the samples might have been assigned to the wrong cluster.

Proposed Evaluation Metrics 3/3

For this challenge:

1. **Compute metrics to evaluate supervised clustering**: For clustering consistency use Silhouette score, for right increment classification of each cluster use purity.

2. **Normalize the metrics between 0 and 1**: Ensure that the metrics for both tasks are normalized, so that they fall between 0 and 1. This step is important to ensure that the metrics are comparable.

3. **Compute the final metric**: calculate the average of the two normalized metrics and subtract a penalty term equal to *0.05* times the numbers of clusters to obtain an overall assessment.

Analysis Results

A fundamental part of a data scientist's work is the ability to narrate and describe the results (storytelling), as it provides crucial insights for making informed decisions, personalizing business strategies and enhancing communication with customers.

Visualization of Features Distribution for Clusters and Interpretation of Results

Provide representative graphs for the clusters, in order to visualize the distribution of features within the different clusters. This analysis will be useful in understanding how the various characteristics are distributed among the various groups identified by clustering.
Additionally, interpret and validate the results obtained.

Example of the task