

# Sistema di Donazione del sangue

Alessandro Lucci - Matricola 1102694

## 1. Introduzione

Il database si compone di diverse tabelle ciascuna delle quali rappresenta un' entità specifica. Le entità scelte per la creazione di questo database sono: Medico, Donatore, Donazione, Centro Di raccolta, Magazzino, Test con le entità figlie Test Sangue intero e Test Infezioni Virali.

## 2. Creazione del Database

### 2.1 Spiegazione delle cardinalità

Le cardinali usate nel sistema sono le seguenti:

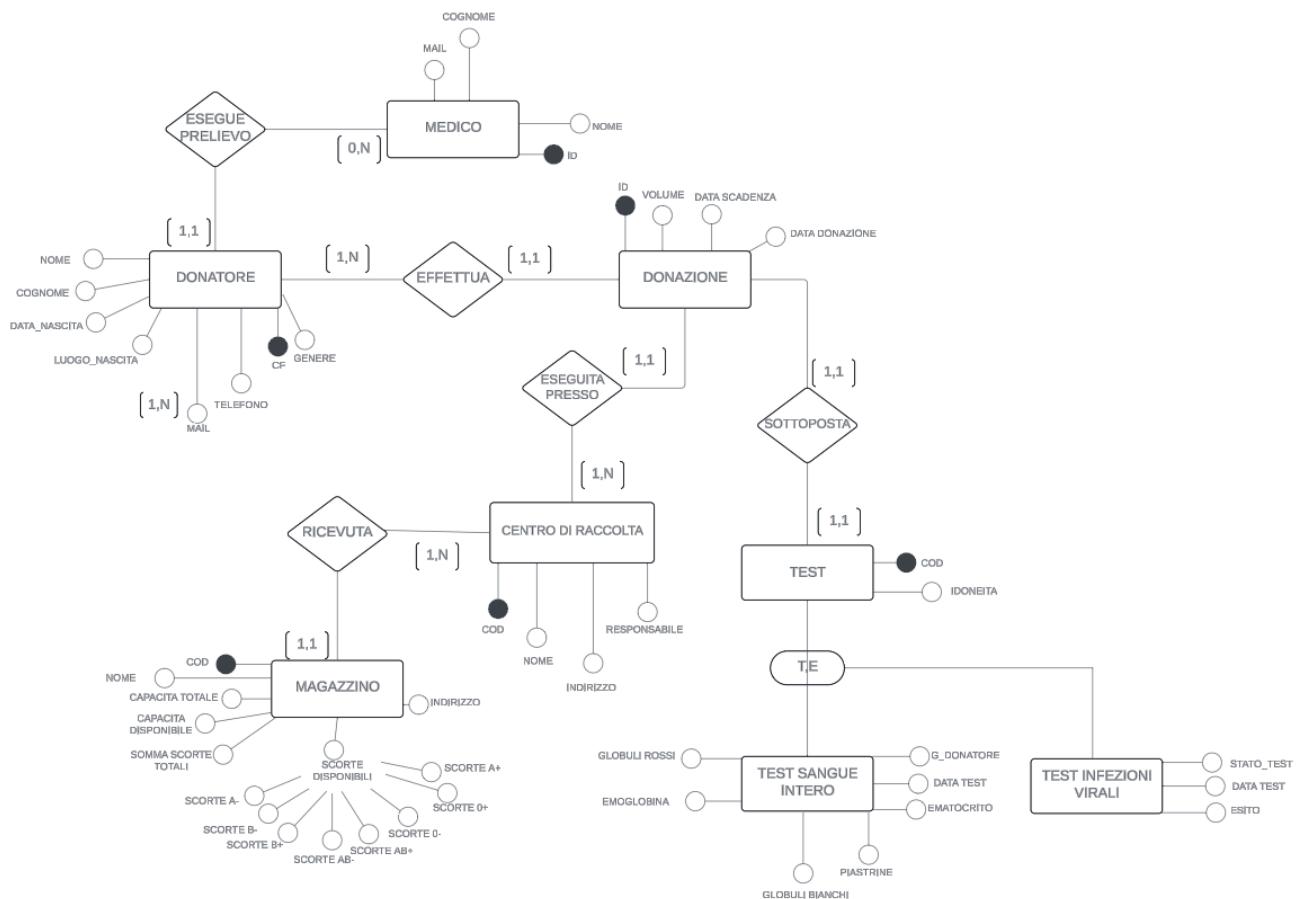
- A. **Medico - Donatore (0-N)** Un medico non ha bisogno di essere associato a un donatore per e può avere più donatori
- B. **Donatore - Medico (1-1)** Un donatore è seguito da un solo medico, non può essere seguito da più medici
- C. **Donatore- Donazione (1-N)** Un donatore deve fare una donazione per essere un donatore ma ne può fare più di una
- D. **Donazione- Donatore (1-1)** Una donazione ha un solo donatore
- E. **Donazione - Centro di Raccolta (1-1)** La singola donazione deve essere eseguita presso uno centro di raccolta e in massimo 1.
- F. **Centro di Raccolta - Magazzino (1-N)** Un centro di raccolta deve avere un magazzino ma ne può avere più di uno, mentre il singolo magazzino è collegato a un singolo centro di raccolta

**G. Donazione - Test (1-1)** Una donazione deve essere sottoposta a dei test e ogni donazione è associata a un singolo set di test che ne include vari test, ma la relazione tra le entità è unica.

**H. Test - Donazione (1-1)** Ogni test è associato a una singola donazione.

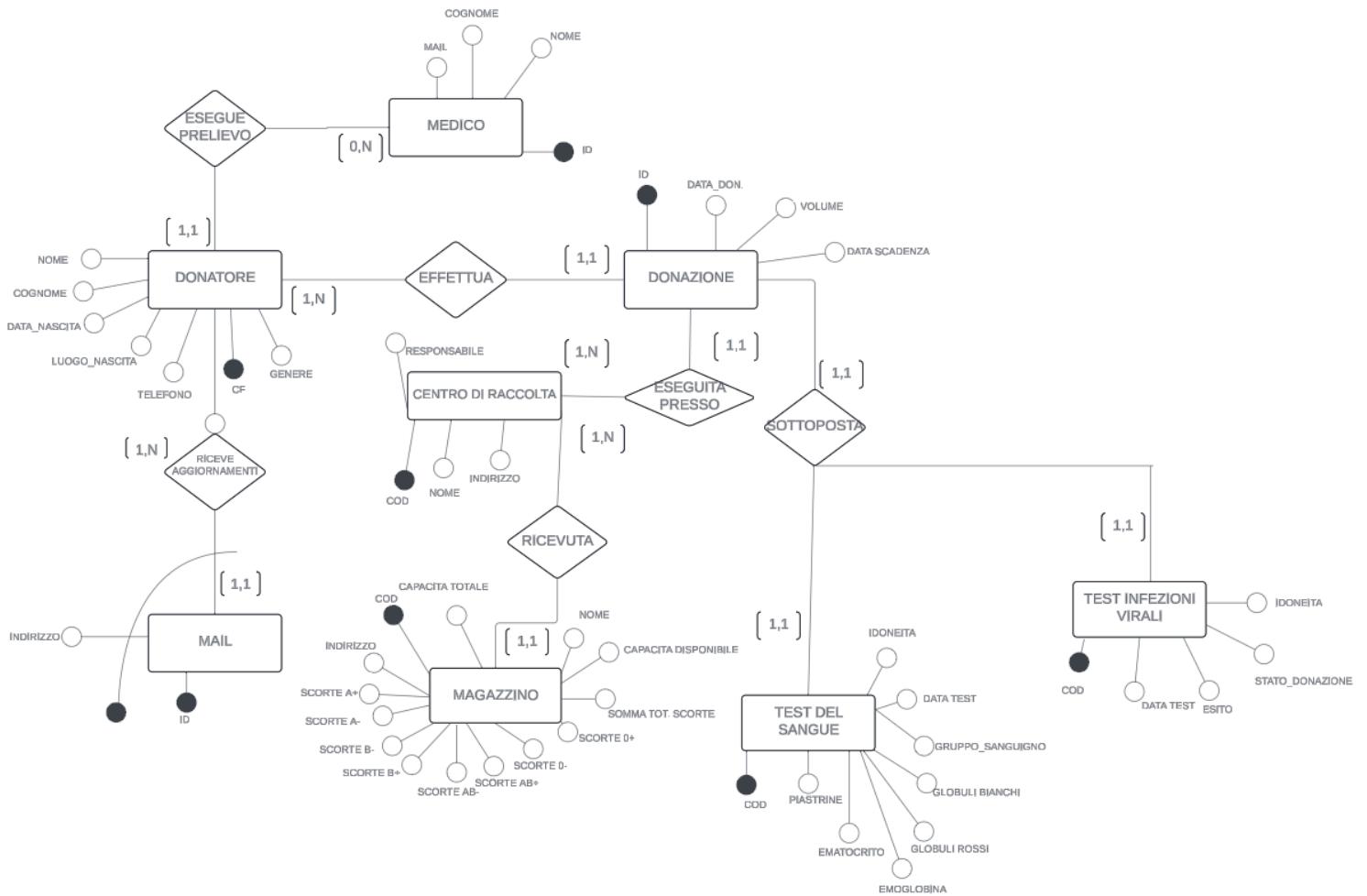
## 2.2 Schema ER non ristrutturato

Qui riportato lo schema ER non ristrutturato, l'**attributo multivalore** scelto è la mail con cui il donatore riceve aggiornamenti e l'**attributo composto** è la disponibilità di scorte divise per tipo.



### 2.3. Schema ER Ristrutturato

Qui riportato lo schema Ristrutturato con la creazione di una nuova entità **mail**, e l'eliminazione dell'attributo composto Scorte Disponibili del magazzino in favore di un insieme di **scorte divise** per tipo. Inoltre, la gerarchia dei test ha visto l'eliminazione del **padre** (Test), essendo **Totale** ed **Esclusiva**, i suoi attributi sono passati ai figli



## 2.4. Modello Logico dello schema ristrutturato

**Medico** (**Id**, Nome, Cognome, Mail)

**Donatore** (**CF**, Nome, Cognome, Genere, DataNascita, LuogoNascita, **Id\_Medico**)

**Mail** (**Id\_Mail**, Indirizzo, **CF\_Donatore**)

**Donazione** (**Cod**, Data, Volume, DataScadenza, **IdMagazzino**, **CF\_Donatore**,  
**Cod\_TestInfezioni**, **Cod\_TestSangue**, **Cod\_CentroRaccolta**)

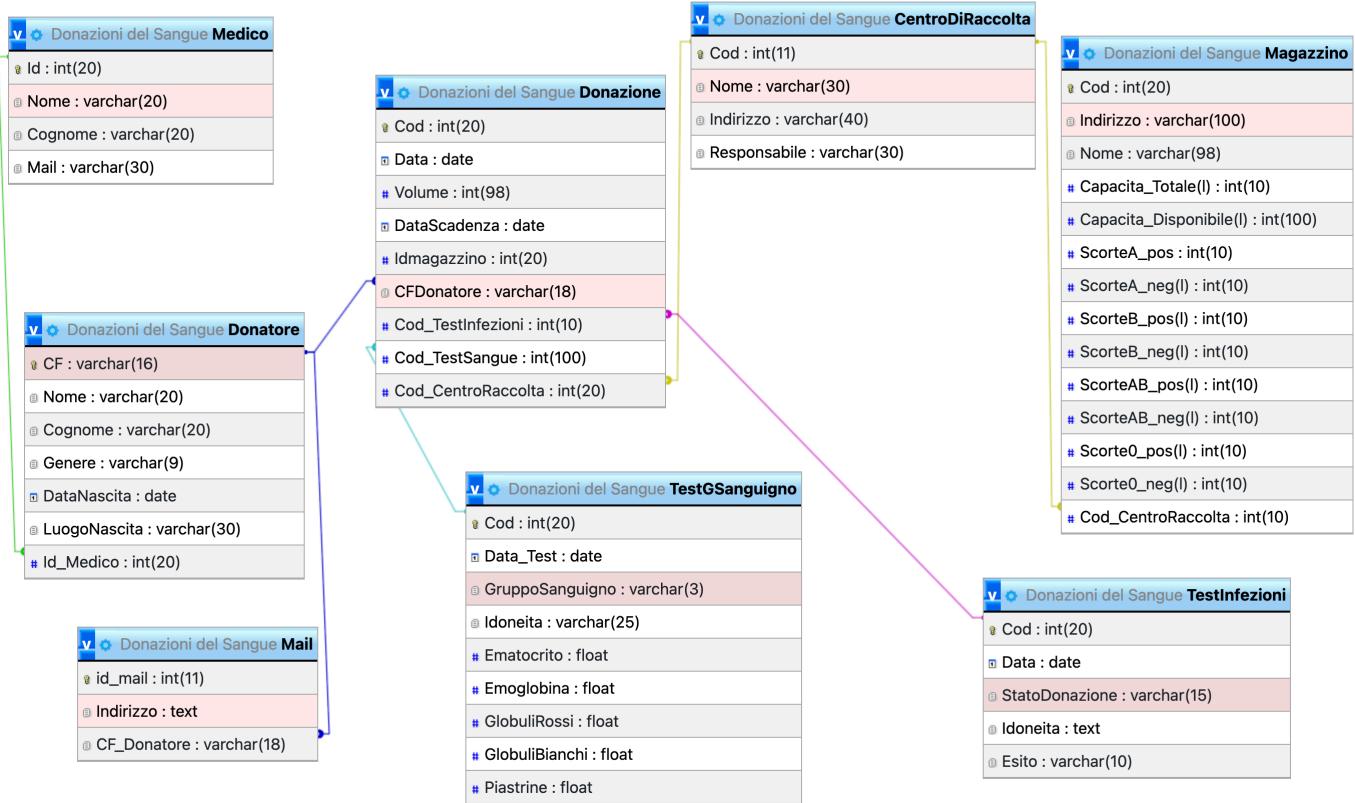
**TestGSanguigno** (**Cod**, Data\_Test, GruppoSanguigno, Idoneità, Ematocrito, Emoglobina, GlobuliRossi, GlobuliBianchi, Piastrine)

**TestInfezioni** (**Cod**, Data, StatoDonazione, Idoneità, Esito)

**CentroDiRaccolta** (**Cod**, Nome, Indirizzo, Responsabile)

**Magazzino** (**Cod**, Indirizzo, Nome, Capacità\_Totale, Capacità\_Disponibile, ScorteA\_pos, ScorteA\_neg, ScorteB\_pos, ScorteB\_neg, ScorteAB\_pos, ScorteAB\_neg, Scorte0\_pos, Scorte0\_neg, **Cod\_CentroRaccolta**)

## 2.5 Schermata delle Relazioni



## 2.6 Query sul Database

L'obiettivo di questa query è identificare il medico che ha eseguito il maggior numero di donazioni associate a donatori.

```

SELECT Medico.Nome, Medico.Cognome, COUNT(Donazione.Cod) AS numero_donazioni
FROM Medico
JOIN Donatore ON Medico.Id = Donatore.Id_Medico
JOIN Donazione ON Donatore.CF = Donazione.CFDonatore
GROUP BY Medico.Nome, Medico.Cognome
ORDER BY numero_donazioni DESC
LIMIT 1;
    
```

Nome	Cognome	numero_donazioni
Kala	Doohey	25

L'obiettivo di questa query è identificare i donatori cui sangue è compatibile con il gruppo 0+, ovvero tutti quelli del gruppo 0.

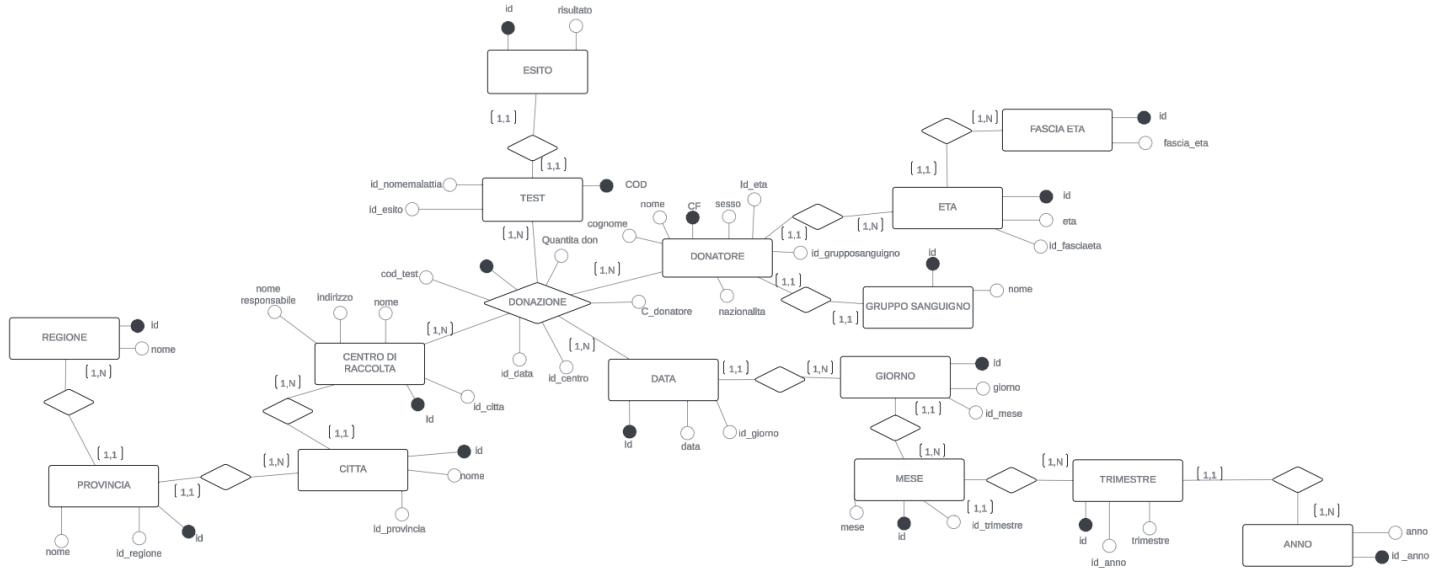
```
SELECT Donatore.Nome, Donatore.Cognome, TestGSanguigno.GrupoSanguigno AS  
GruppoSanguigno_Donazione  
FROM Donatore  
JOIN Donazione ON Donazione.CFDonatore = Donatore.CF  
JOIN TestGSanguigno ON TestGSanguigno.Cod = Donazione.Cod_TestSangue  
WHERE TestGSanguigno.GrupoSanguigno LIKE '0%';
```

Nome	Cognome	GruppoSanguigno_Donazione
Chiara	Greco	0+
Alessandro	Romano	0-
Elisa	Ricci	0-
Giovanni	Colombo	0+
Francesca	Romano	0-
Giovanni	Russo	0+
Francesca	Ferrari	0+
Matteo	Marino	0+
Chiara	Colombo	0-
Martina	Russo	0+
Martina	Marino	0+

### 3. Creazione del Data Warehouse

Il seguente Data Warehouse è stato creato con le seguenti finalità: tracciamento della distribuzione geografica e analisi temporali delle donazioni, monitoraggio dei test eseguiti ed i loro esiti e analisi dettagliate sulla popolazione dei donatori.

### 3.1 Modello a fiocco di neve



### 3.2 Modello Logico

1. **Donazione** (**Id\_donazione**, QuantitaDonata, NumeroDonazioni, **id\_donatore**, **id\_data**, **id\_centro**, **cod\_test**)
2. **Donatore**(**CF**, nome, cognome, sesso, nazionalità, **id\_grupposanguigno**, **id\_eta**)
3. **Eta**(**id\_eta**, eta, **id\_fasciaeta**)
4. **Fascia Eta**(**id\_fasciaeta**, fascia\_eta)
5. **Gruppo\_Sanguigno** (**id\_gs**, nome)
6. **Data**(**id\_data**, data, **id\_giorno**)
7. **Giorno**(**id\_giorno**, giorno, **id\_mese**)

8. **Mese**(**id\_mese**, mese, **id\_trimestre**)

9. **Trimestre**(**id\_trimestre**, trimestre, **id\_anno**)

10. **Anno** (**id\_anno**, anno)

11. **Test** (**COD\_test**, nome)

12. **Esito** (**id\_esito**, risultato)

13. **CentroRaccolta** (**id\_centro**, nome, indirizzo, responsabile, **id\_citta**)

14. **Citta**(**id\_citta**, nome, **id\_provincia**)

15. **Provincia** (**id\_provincia**, nome, **id\_regione**)

16. **Regione**(**id\_regione**, nome)

### 3.3 Operazioni OLAP

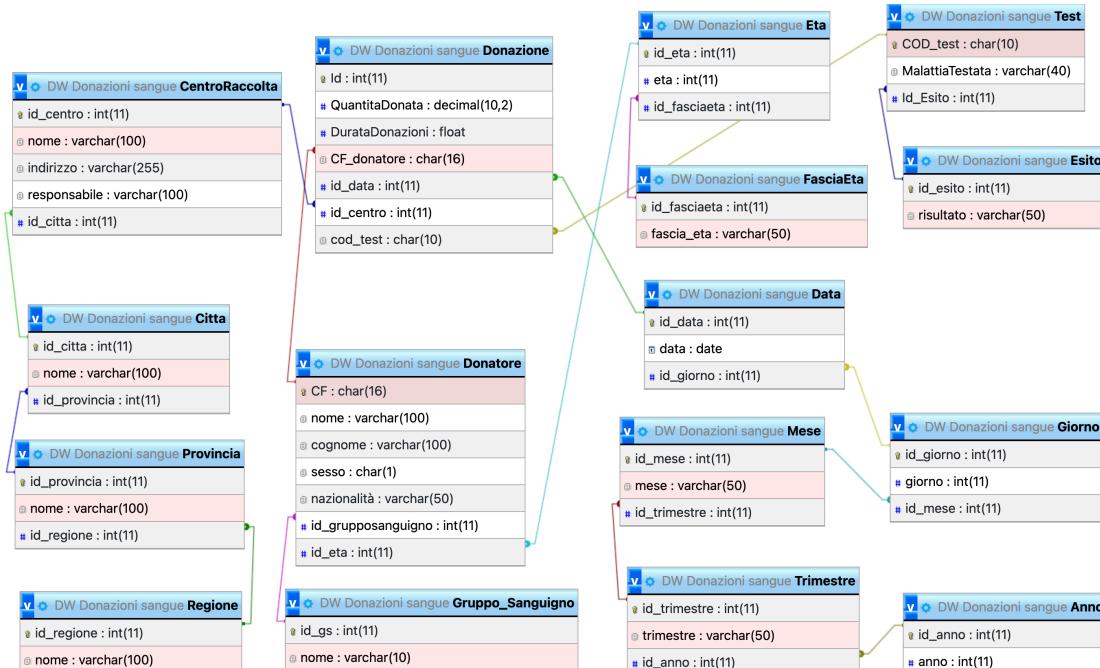
Con questa operazione Slice otteniamo il totale delle donazioni per ogni fascia d'eta.

<b>Fascia_Età</b>	<b>COUNT di Id_Donazione</b>
18-24 anni	6
25-34 anni	16
35-44 anni	10
45-54 anni	20
55-60 anni	18
<b>Totale generale</b>	<b>70</b>

Con questa operazione di Drill-down vogliamo passare da un livello Regione a un livello Città, per ottenere un analisi granulare dei dati delle donazioni.

Regione	Provincia	Città	COUNT di Id_Donazione
- Lazio	- Frosinone	Frosinone	4
	<b>Totale Frosinone</b>		<b>4</b>
	- Latina	Aprilia	3
		Latina	5
	<b>Totale Latina</b>		<b>8</b>
	- Rieti	Rieti	4
	<b>Totale Rieti</b>		<b>4</b>
	- Roma	Roma	10
	<b>Totale Roma</b>		<b>10</b>
	- Viterbo	Viterbo	4
	<b>Totale Viterbo</b>		<b>4</b>
<b>Totale Lazio</b>			<b>30</b>
<b>Totale generale</b>			<b>30</b>

### 3.4. Schermata delle relazioni

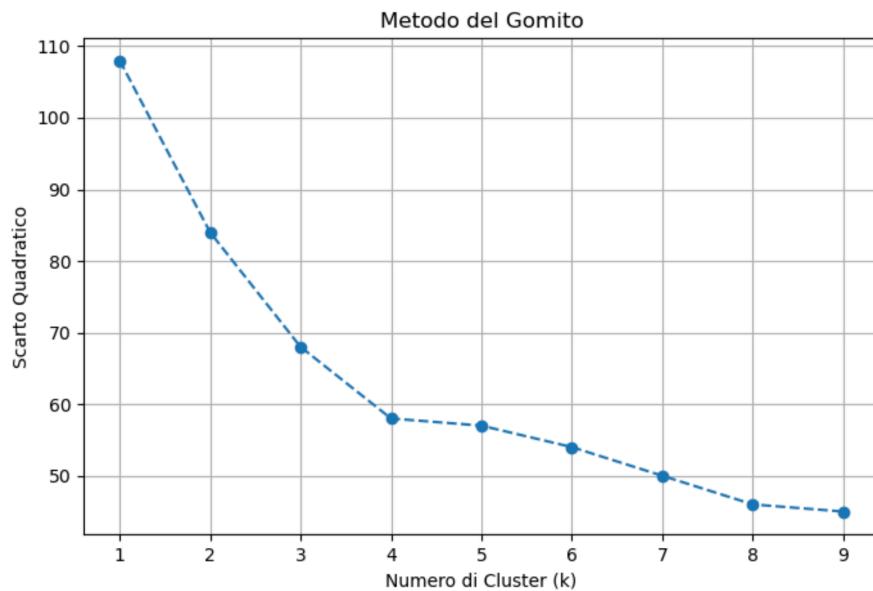


## 4. Data Mining

**Algoritmo Utilizzato:** K-means

**Misura di distanza :** Euclidea

**Numero di Cluster(k):** 4, per individuare il numero di cluster ottimali è stato usato il metodo del gomito con i seguenti dati (108:1, 84:2, 68:3, 58:4, 57:5, 54:6, 50:7, 46



**Seed utilizzato:** 10

**Scarto Quadratico:** 58

## 4.1 Analisi del cluster

```
== Clustering model (full training set) ==

kMeans
=====

Number of iterations: 4
Within cluster sum of squared errors: 58.0

Initial starting points (random):
Cluster 0: B+, '35-44 anni'
Cluster 1: B+, '55-60 anni'
Cluster 2: AB+, '45-54 anni'
Cluster 3: O+, '55-60 anni'

Missing values globally replaced with mean/mode

Final cluster centroids:
Attribute      Full Data          Cluster#
                  (70.0)           0       1       2       3
                                         (27.0)  (19.0)  (17.0)  (7.0)
=====
Gruppo_Sanguigno   A-            B+            A+            A-            O+
AB+                8.0 ( 11%) 2.0 (  7%) 2.0 ( 10%) 3.0 ( 17%) 1.0 ( 14%)
A+                10.0 ( 14%) 3.0 ( 11%) 7.0 ( 36%) 0.0 (  0%) 0.0 (  0%)
O+                5.0 (  7%) 1.0 (  3%) 1.0 (  5%) 0.0 (  0%) 3.0 ( 42%)
AB-                7.0 ( 10%) 1.0 (  3%) 3.0 ( 15%) 3.0 ( 17%) 0.0 (  0%)
O-                10.0 ( 14%) 5.0 ( 18%) 1.0 (  5%) 2.0 ( 11%) 2.0 ( 28%)
B+                10.0 ( 14%) 10.0 ( 37%) 0.0 (  0%) 0.0 (  0%) 0.0 (  0%)
A-                12.0 ( 17%) 2.0 (  7%) 4.0 ( 21%) 6.0 ( 35%) 0.0 (  0%)
B-                8.0 ( 11%) 3.0 ( 11%) 1.0 (  5%) 3.0 ( 17%) 1.0 ( 14%)
Fascia_Età        45-54 anni 25-34 anni 55-60 anni 45-54 anni 35-44 anni
45-54 anni        20.0 ( 28%) 3.0 ( 11%) 1.0 (  5%) 16.0 ( 94%) 0.0 (  0%)
55-60 anni        18.0 ( 25%) 1.0 (  3%) 17.0 ( 89%) 0.0 (  0%) 0.0 (  0%)
25-34 anni        16.0 ( 22%) 16.0 ( 59%) 0.0 (  0%) 0.0 (  0%) 0.0 (  0%)
35-44 anni        10.0 ( 14%) 2.0 (  7%) 0.0 (  0%) 1.0 (  5%) 7.0 (100%)
18-24 anni        6.0 (  8%) 5.0 ( 18%) 1.0 (  5%) 0.0 (  0%) 0.0 (  0%)
```

L'analisi è stata effettuata tramite l'algoritmo k means con k=4, l'obiettivo è identificare dei gruppi di donatore in base alle caratteristiche Gruppo Sanguigno e Fascia d'età, così da comprendere il profilo dei donatori e fare delle strategie di comunicazione mirate a loro.

L'analisi ha generato 4 cluster

- **Cluster 0.** Fascia di età: 25-34 anni GruppoS: B+

Questo cluster rappresenta giovani adulti con il gruppo sanguigno B+, un gruppo abbastanza comune ma fondamentale.

- **Cluster 1** Fascia di età: 55-60 anni GruppoS: A+

Questo cluster rappresenta donatori adulti con il gruppo sanguigno A+, in questo cluster la fascia d'eta 55-60 è molto predominante e omogenea , circa l'89 %. Inoltre ci segnala la necessità di ringiovanire questa categoria , poiché l'eta avanzata ne riduce la futura disponibilità.

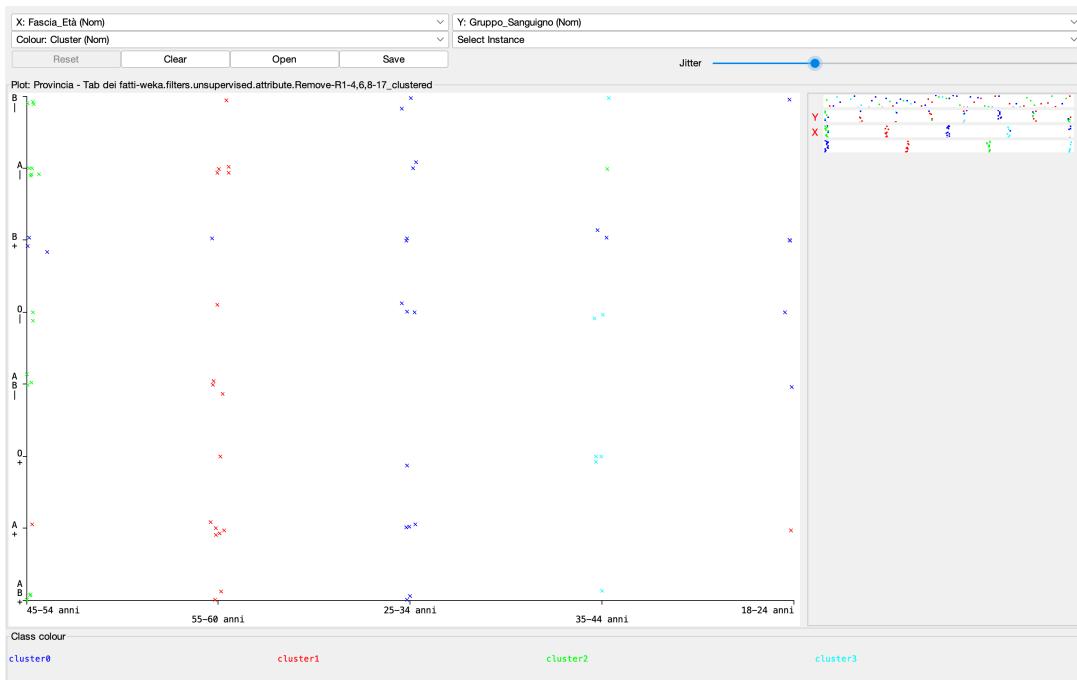
- **Cluster 2** Fascia di eta: 45-54 anni GruppoS: A-

Questo cluster rappresenta dei donatori di mezza eta, ed è particolarmente significativo perché include due gruppi molto rari(A-, AB-). Inoltre questi donatori hanno ancora un minimo di 5 e un massimo di 15 anni di donazioni possibili.

- **Cluster 3** Fascia di eta: 35-44 anni GruppoS: 0+

Questo cluster rappresenta dei donatori di prima mezza eta con il gruppo predominante 0+, uno dei più comuni ma importante per la disponibilità del sangue. Inoltre rappresenta il cluster più piccolo ma anche uno dei più giovani , cosa che da molta rilevanza. Possiamo affermare che il cluster 3 sia il più **omogeneo** tra i 4, grazie alla predominanza del gruppo 0+ e alla quasi totale uniformità della fascia eta 35-44 anni.

L'analisi dimostra che il cluster più importante è il **cluster 2**, in quanto composto principalmente da A- e AB- dei gruppi rari e risorsa critica per le situazioni di emergenza, inoltre bisogna rimarcare l'importanza del cluster 3, composta da donatori di una fascia d'eta stabile e con il gruppo 0+, il più richiesto dalle donazioni.



Le variabili del seguente grafico sono:

**Asse X:** Fascia eta suddivisa in diversi intervalli

**Asse Y:** Gruppo sanguigno suddiviso in categorie

Ogni cluster tende a concentrarsi su degli intervalli di specifiche. Per esempio;

- **Cluster 0 (blu)** si distribuisce soprattutto nella fascia **25-34 anni**, ma mantiene una buona distribuzione anche nelle altre fasce d'eta.
- **Cluster 1 (rosso)** si distribuisce nella fascia **55-60 anni**.
- **Cluster 2 (verde)** si distribuisce nella fascia **45-54 anni**.
- **Cluster 3 (celeste)** si distribuisce nella fascia **35-44 anni**.

Non sono presenti delle sovrapposizioni evidenti tra i cluster, evidenziando una buona separazione tra i gruppi.

### 4.3 Albero decisionale

**Variabili di input:** Sesso\_Donatore, Gruppo Sanguigno

**Variabili di output:** fascia di età

**Tecnica di valutazione:** Cross Validation (10 fold)

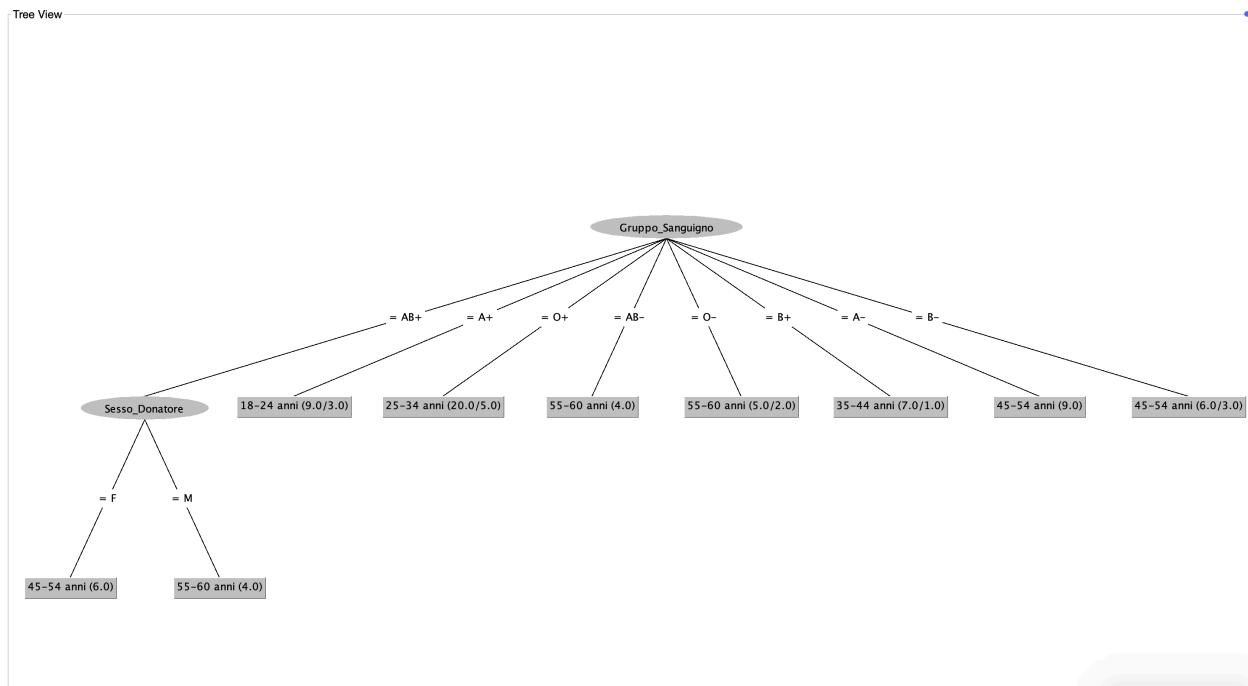
**Precision:** 0,814

**Recall:** 0,800

Correctly Classified Instances	56	80	%						
Incorrectly Classified Instances	14	20	%						
Kappa statistic	0.7422								
Mean absolute error	0.1184								
Root mean squared error	0.2433								
Relative absolute error	38.2481 %								
Root relative squared error	61.9142 %								
Total Number of Instances	70								
<b>==== Detailed Accuracy By Class ====</b>									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,900	0,060	0,857	0,900	0,878	0,828	0,983	0,945	45-54 anni
	0,611	0,038	0,846	0,611	0,710	0,644	0,879	0,769	55-60 anni
	0,938	0,093	0,750	0,938	0,833	0,785	0,948	0,743	25-34 anni
	0,600	0,017	0,857	0,600	0,706	0,680	0,914	0,670	35-44 anni
	1,000	0,047	0,667	1,000	0,800	0,797	0,977	0,667	18-24 anni
Weighted Avg.	0,800	0,055	0,814	0,800	0,793	0,747	0,938	0,790	
<b>==== Confusion Matrix ====</b>									
	a	b	c	d	e	<-- classified as			
18	0	0	0	2		a = 45-54 anni			
2	11	3	1	1		b = 55-60 anni			
0	1	15	0	0		c = 25-34 anni			
1	1	2	6	0		d = 35-44 anni			
0	0	0	0	6		e = 18-24 anni			

La matrice di confusione ha una dimensione di 5x5, con le seguenti 5 classi:

- **Fascia d'eta 45-54:** composta da 18 valori classificati correttamente, ma 2 falsi positivi in e, e 3 falsi negativi suddivisi in (b=2, d=1)
- **Fascia d'eta 55-60:** composta da 11 valori classificati correttamente, ma una totalità di 5 falsi positivi così suddivisi (a=2, c=3, d=1, e=1) e 2 falsi negativi suddivisi in (c=1, d=1)
- **Fascia d'eta 25-34:** composta da 15 valori classificati correttamente, ma 1 falso positivo in b e 5 falsi negativi suddivisi in (b=3, d=2)
- **Fascia d'eta 35-44:** composta da 6 valori classificati correttamente, ma 4 falsi positivi così suddivisi (a=1, b=1 ,c=2) e 1 falso negativo in b.
- **Fascia d'eta 18-24:** composta da 6 valori classificati correttamente, ma 3 falsi negativi suddivisi così(a=2, b=1)



Il seguente albero rappresenta un modello che utilizza due variabili principali ovvero “Sesso Donatore” e “Gruppo Sanguigno”, per predire la fascia d’eta dei donatori. Il modello inizia con una distinzione basata sul gruppo sanguigno, solo nel gruppo AB+ viene fatta una distinzione sul sesso del donatore, le foglie dell’albero contengono le fasce di età predette.

Le regole derivate da questo albero sono le seguenti:

1. Se Sesso\_Donatore è **F** e il gruppo sanguigno è **AB+**, allora la fascia di età è **45-54 anni**.
2. Se Sesso\_Donatore è **M** e il gruppo sanguigno è **AB+**, allora la fascia di età è **55-60 anni**.
3. Se il gruppo sanguigno è **A+**, allora la fascia di età è **25-34 anni**.
4. Se il gruppo sanguigno è **O+**, allora la fascia di età è **55-60 anni**.
5. Se il gruppo sanguigno è **AB-**, allora la fascia di età è **55-60 anni**.
6. Se il gruppo sanguigno è **O-**, allora la fascia di età è **35-44 anni**.

7. Se il gruppo sanguigno è **B+**, allora la fascia di età è **45-54 anni**.

8. Se il gruppo sanguigno è **B-**, allora la fascia di età è **45-54 anni**.

## 4.5. Modello Naive Bayes

== Summary ==										
Correctly Classified Instances			59	84.2857 %						
Incorrectly Classified Instances			11	15.7143 %						
Kappa statistic			0.7964							
Mean absolute error			0.1515							
Root mean squared error			0.2421							
Relative absolute error			48.9232 %							
Root relative squared error			61.595 %							
Total Number of Instances			70							
== Detailed Accuracy By Class ==										
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
0,800	0,020	0,941	0,800	0,865	0,822	0,956	0,917	0,917	45-54 anni	
0,889	0,096	0,762	0,889	0,821	0,756	0,923	0,777	0,777	55-60 anni	
0,938	0,056	0,833	0,938	0,882	0,847	0,969	0,827	0,827	25-34 anni	
0,600	0,017	0,857	0,600	0,706	0,680	0,891	0,642	0,642	35-44 anni	
1,000	0,016	0,857	1,000	0,923	0,919	0,992	0,857	0,857	18-24 anni	
Weighted Avg.	0,843	0,047	0,851	0,843	0,840	0,799	0,944	0,816		
== Confusion Matrix ==										
a	b	c	d	e	<-- classified as					
16	3	0	0	1		a = 45-54 anni				
0	16	1	1	0		b = 55-60 anni				
0	1	15	0	0		c = 25-34 anni				
1	1	2	6	0		d = 35-44 anni				
0	0	0	0	6		e = 18-24 anni				

Mettendo a confronto i risultati ottenuti dal modello **J48** di prima e il modello **Naive bayes** possiamo notare che:

Il modello Naive Bayes ha un accuratezza di classificazione delle istanze migliore con un tasso del **84.28%** rispetto al **80%** del modello J48.

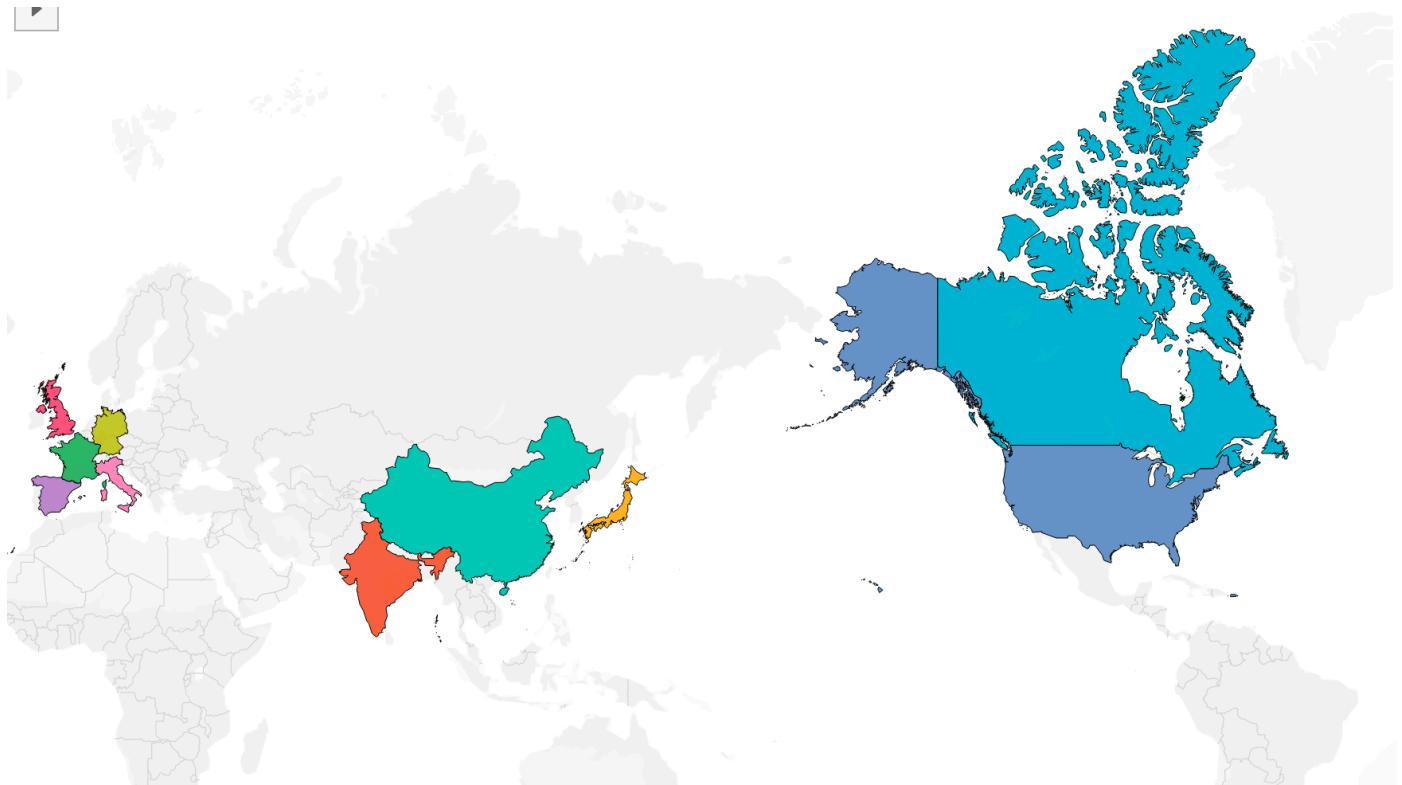
Il modello J48 ha una media errori e uno squarto quadratico medio minori rispetto a quelli del naive Bayes (**0.1184, 0.2421**) (**0.2433, 0.4923**)

Per quanto riguarda la precision, recall e la f-measure, il modello naive bayse ottiene dei risultati migliori in media(P=**0,814,0,851**) (R=**0,800,0,843**) (F=**0,793,0,840**).

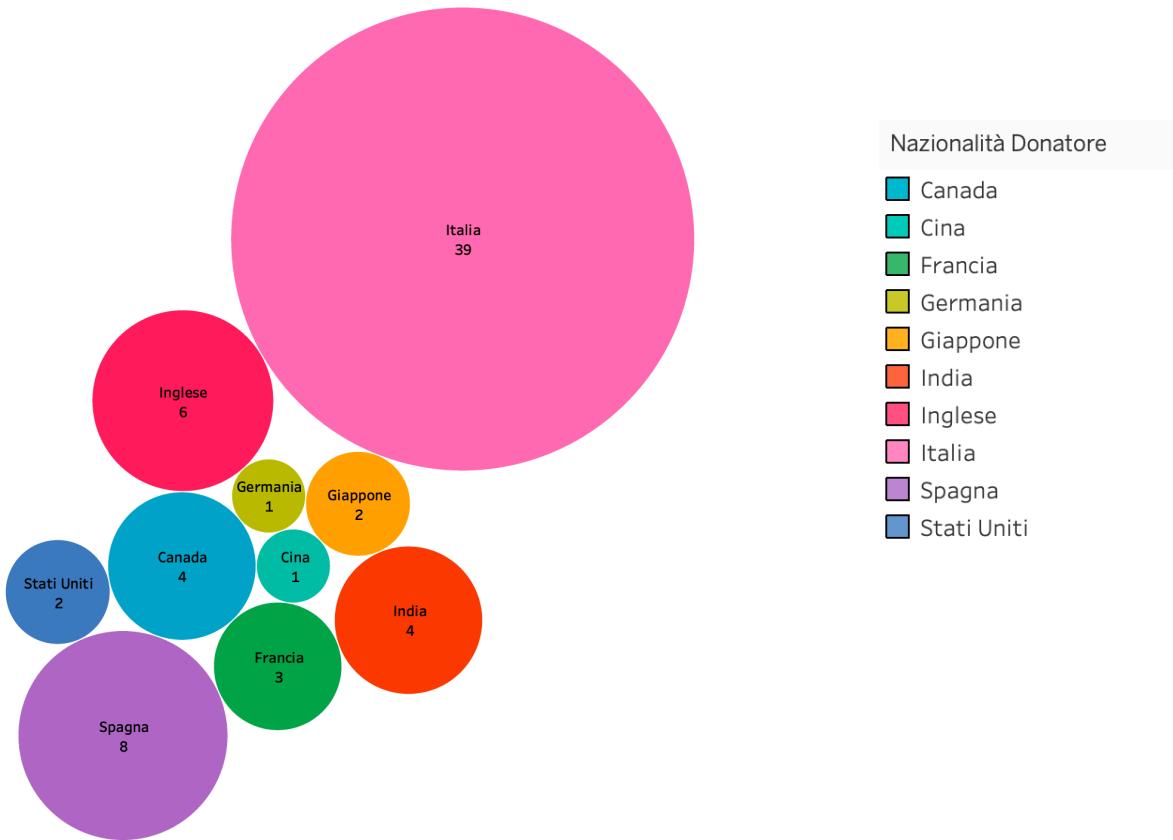
Confrontando a matrice di confusione del naive Bayes con quella J48 appare evidente che il modello naive bayes abbia la miglior predizione nella fascia d'eta b=55-60 anni e c=25-34, grazie alla minore presenza di falsi negativi e falsi positivi.

## 5. Info Visualization

Ho usato Tableau per creare delle visualizzazioni utili a comprendere meglio i dati. Ho realizzato una mappa che mostra la distribuzione delle nazionalità dei donatori e un grafico a bolle per evidenziare la dimensione di ogni nazionalità. Poi, ho aggiunto un grafico a barre che mette in evidenza i gruppi sanguigni più donati, integrandolo con una linea di media per rendere i dati più facili da interpretare.



**Grafico a bolle che rappresenta la dimensione relativa di ogni nazionalità**



**Grafico a barre che rappresenta i gruppi sanguigni più donati con una linea di media.**

