Manuale d'uso di Pentaho per analisi OLAP

in Immunologia

Denitto Matteo

INDICE:

1. INTRODUZIONE

Con questo manuale si vuole curare e analizzare in dettaglio l'analisi dei dati e la creazione di grafici attraverso l'utilizzo di Pentaho (software OpenSource di Business Intelligence), pertanto verranno affrontati solamente gli aspetti di questo software che prendono parte a questi processi.

2. COME INIZIARE

Il modo più facile per accedere alla console di Pentaho è collegandosi all'URL del server attraverso un browser. Nel caso si tratti di un accesso in locale l'URL è: http://localhost/8080/. In Figura 1 si possono notare i primi passi da compiere che permettono l'accesso all'interfaccia vera e propria di Pentaho (in rosso sono evidenziati i bottoni sui quali cliccare).

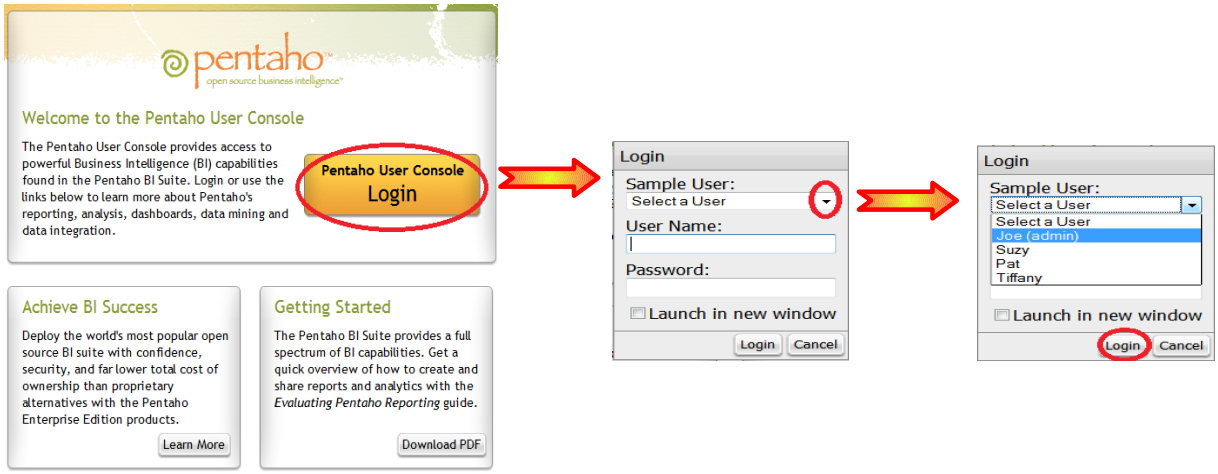


Figura 1: procedura di autenticazione per Pentaho BI Suite

***N.B.** La selezione dell'account utente da utilizzare (Joe, Suzy, Pat, Tiffany) è ininfluente ai fini di questo manuale; sarà comunque possibile utilizzare il proprio account una volta che Pentaho verrà installato e reso funzionante.*

Cliccando su "Login" appare l'interfaccia di Pentaho che permette di selezionare la tipologia di lavoro che si vuole svolgere, nel nostro caso è sufficiente cliccare su "New Analysis View" per poter aver accesso alla lista degli "Schema" salvati.
Uno Schema rappresenta l'insieme dei dati sui quali è permesso lavorare; tra tutti quelli presenti occorre selezionare lo Schema di interesse, in questo caso ImmunoRRR.

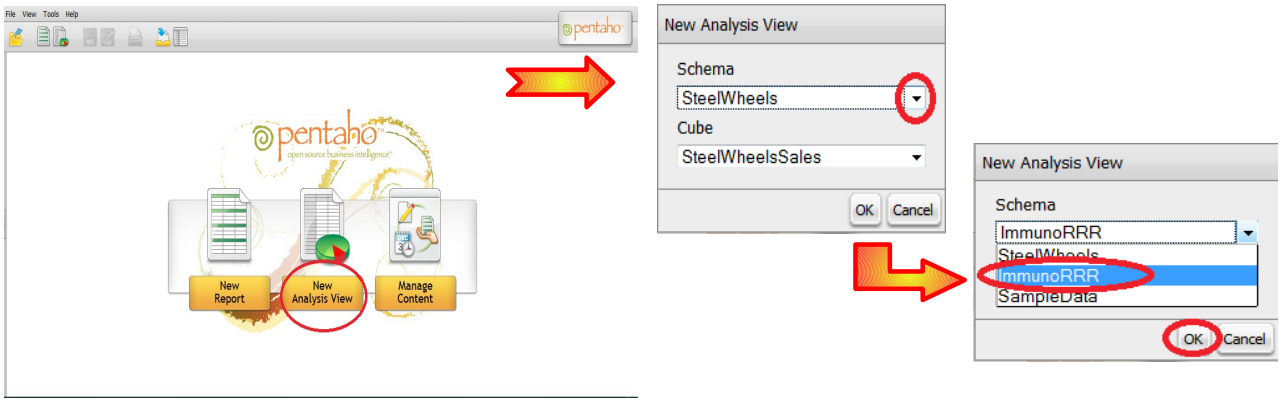


Figura 2: Accesso alla sezione di Pentaho dedicata alle analisi

Cliccando su "OK" si presenta l'interfaccia che permette di analizzare i dati ed effettuare i grafici sugli stessi.

3. COME UTILIZZARE LA TABELLA

In Figura 3 viene mostra come si presenta la pagina una volta effettuato l'accesso ed essere entrati nella sezione relativa alle analisi.

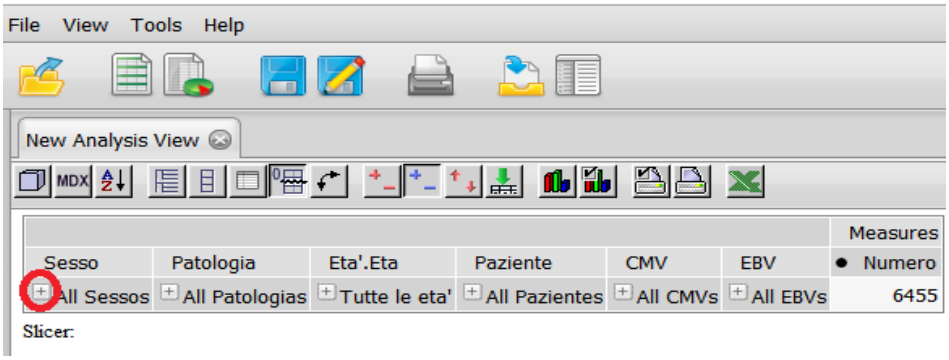


Figura 3: Tabella contenente i dati

Come si può facilmente intuire dalla schermata, ci sono tasti intuitivi e di uso comune che permettono di aprire file esistenti, salvare il lavoro corrente e stampare il grafico che si sta visualizzando.
Gli altri tasti più in basso servono a personalizzare il lavoro che si sta svolgendo, quando saranno utilizzati ne verrà spiegato anche il funzionamento.
Per il momento concentriamoci sulla tabella.
La tabella è divisa in categorie sia per quel che riguarda le righe, sia per le colonne; nel nostro caso le righe sono suddivise per "Sesso", "Patologia", "Età", "Paziente", "CMV" (Citomegalovirus) e "EBV" (Epstein e Barr virus) mentre per il momento l'unica colonna esistente è rappresentata dal campo Measures (in seguito vedremo come modificare questo campo, per il momento viene utilizzato "Numero" perchè più utile ai fini esplicativi).
La tabella così come viene mostrata non ha molto interesse, per poter apprezzare la sua utilità occorre modificarla. E' possibile espandere (e comprimere) i vari campi della tabella semplicemente cliccando sui pulsanti che si trovano a fianco del nome. Ad esempio, se si volesse sapere quanti sono in numero i pazienti e differenziarli per sesso basta cliccare sulla croce che sta a fianco del nome sesso per ottenere in Figura 4.

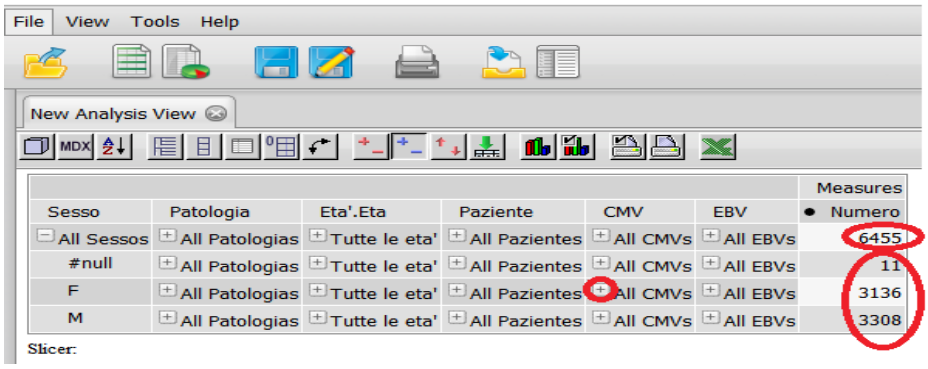
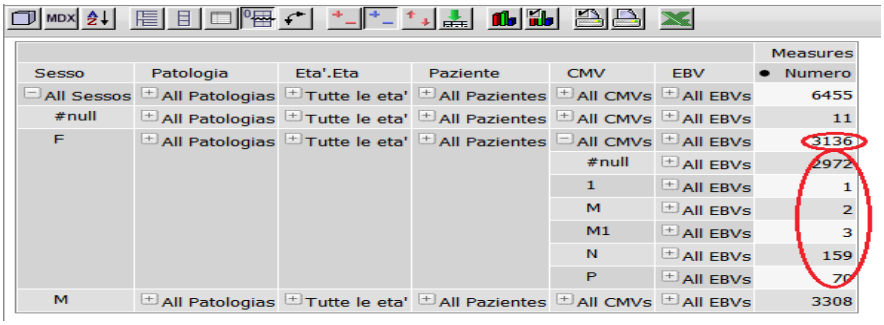


Figura 4: Esempio di tabella in cui è stata espansa una sezione

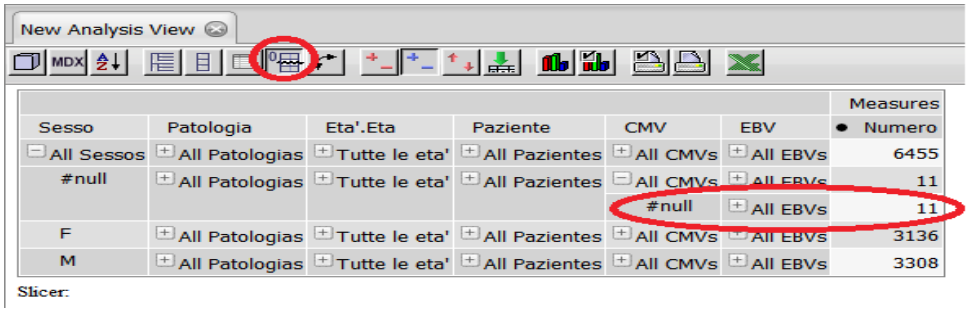
Come si può notare quello che si ottiene è una separazione basata sul sesso; inoltre è importante capire come varia il numero rappresentato nel campo Measures, in questo caso è piuttosto semplice e quello che si ottiene sta a significare che tra tutti i pazienti inseriti (6455) i maschi sono 3308, le femmine 3136 e i "#null" sono 11 (con #null si indica, all'interno di ogni contesto, qualcosa di sconosciuto o di cui non si ha a disposizione quel tipo di dato; qui sta a significare che ci sono 11 persone di cui non si conosce il sesso).
Si può approfondire ulteriormente la ricerca andando, per esempio, ad analizzare i dati dei Citomegalovirus per quel che riguarda le donne.



Sesso	Patologia	Eta'.Eta	Paziente	CMV	EBV	Measures: Numero
All Sessos	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	6455
#null	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	11
F	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	3136
				#null	All EBVs	2972
				1	All EBVs	1
				M	All EBVs	2
				M1	All EBVs	3
				N	All EBVs	159
				P	All EBVs	70
M	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	3308

Figura 5: Esempio di tabella con due sezioni espanse

Nei nostri dati quello che si può notare è che delle 3136 donne, una possiede CMV 1, 2 possiedono M e così via, per 2972 donne non si possiede quest'informazione. Notare che in questo caso se si sommano le diverse categorie di CMV (compresi i null) si ottiene un valore più alto di 3136, questo perchè una persona può essere infettata da più tipi di CMV.
Riproviamo a fare ora la stessa analisi utilizzando le persone di cui non si conosce il sesso.

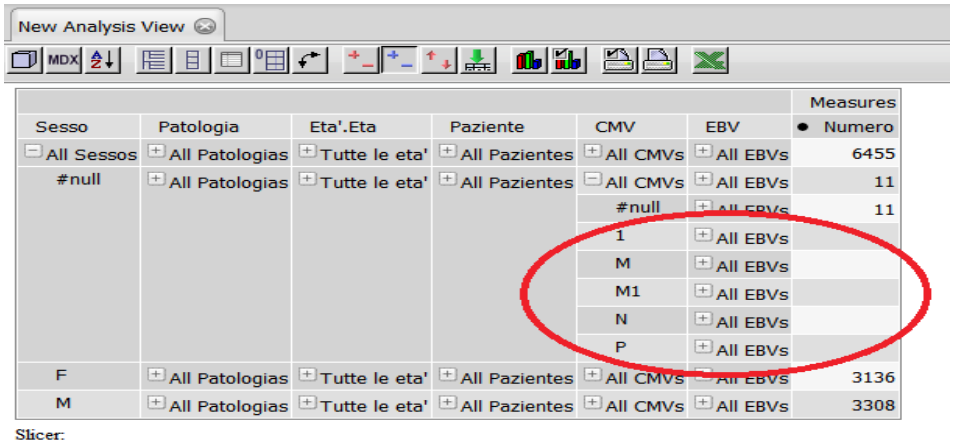


Sesso	Patologia	Eta'.Eta	Paziente	CMV	EBV	Measures: Numero
All Sessos	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	6455
#null	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	11
				#null	All EBVs	11
F	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	3136
M	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	3308

Slicer:

Figura 6: Utilizzo del "Mostra/Nascondi colonne vuote"

Quello che si ottiene è una tabella in cui è presente una sola riga (a differenza delle 6 di quella precedente), questo perchè tutti pazienti di cui non si conosceva il sesso condividono lo stesso status per quel che riguarda CMV (#null). Tuttavia è possibile vedere il risultato nello stesso formato della tabella precedente cliccando sull'icona cerchiata in rosso, questa permette di visualizzare/eliminare le righe vuote presenti nella tabella.



Sesso	Patologia	Eta'.Eta	Paziente	CMV	EBV	Measures: Numero
All Sessos	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	6455
#null	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	11
				#null	All EBVs	11
				1	All EBVs	
				M	All EBVs	
				M1	All EBVs	
				N	All EBVs	
				P	All EBVs	
F	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	3136
M	All Patologias	Tutte le eta'	All Pazientes	All CMVs	All EBVs	3308

Slicer:

Figura 7: Utilizzo del "Mostra/Nascondi colonne vuote"

E' possibile analizzare i dati in questo modo navigando attraverso l'intera tabella, tuttavia è sconsigliato perchè si potrebbe verificare un rallentamento importante nell'esecuzione e perchè visualizzare una tabella di dimensioni elevate può diminuire notevolmente la semplicità di lettura della stessa.
Una volta capito come è strutturata la tabella si può passare all'utilizzo dei grafici disponibili in Pentaho.

4. COME CREARE GRAFICI E UTILIZZARNE LE OPZIONI

Per iniziare a creare dei grafici è sufficiente cliccare sul bottone corrispondente (in Figura 8), quello che viene visualizzato è un grafico che rappresenta tutti i dati senza alcuna categorizzazione e quindi di utilità limitata.

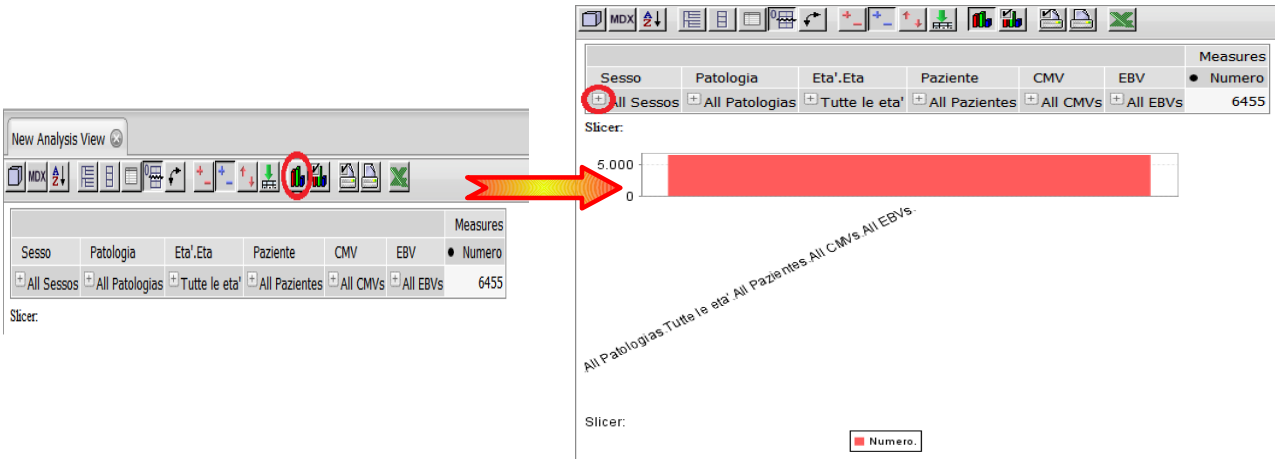


Figura 8: Funzione del tasto "Visualizza Grafici"

E' possibile personalizzare il grafico in due modi: agendo sulla tabella oppure oppure impostando i filtri relativi alle dimensioni e alle misure del "cubo".
In questo capitolo vedremo come viene modificato il grafico quando si agisce sulla tabella come anticipato nel capitolo precedente.
Per cominciare ad interpretare l'interfaccia dove viene rappresentato il grafico conviene crearne uno un po' più interessante; ad esempio cliccando sulla croce vicino a "Sesso" ed espandendo quindi la tabella il grafico viene modificato automaticamente.

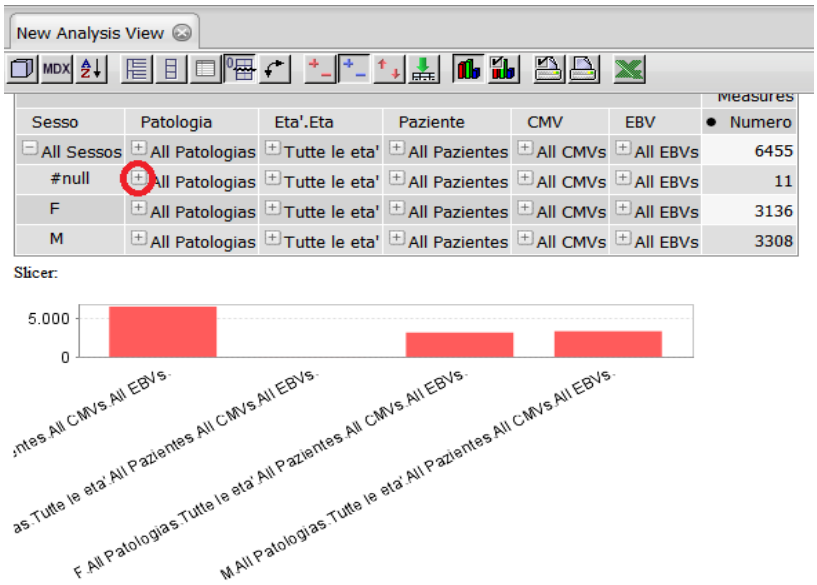


Figura 9: Esempio di istogramma in Pentaho

Il grafico rappresenta di volta in volta tutte le righe presenti nella tabella che si sta visualizzando, in questo caso la prima barra del grafico rappresenta la prima riga della tabella, la seconda barra rappresenta la seconda riga e così via. Sotto ogni barra del grafico è indicato quale parte della tabella rappresenta, prendendo ad esempio la terza si legge che quella parte di grafico descrive i pazienti di sesso femminile

considerando tutte le patologie, tutte le classi di età le altre categorie. Sotto il grafico inoltre è presente anche una legenda che indica che tipo di dato il grafico sta rappresentando (si vede meglio nell'immagine precedente, Figura 9).
Proseguendo con il tutorial, clicchiamo ora nella riga dei "#null" la croce che sta vicino a "All.Patologias" per espandere questa sezione, notiamo che il grafico viene modificato come segue:

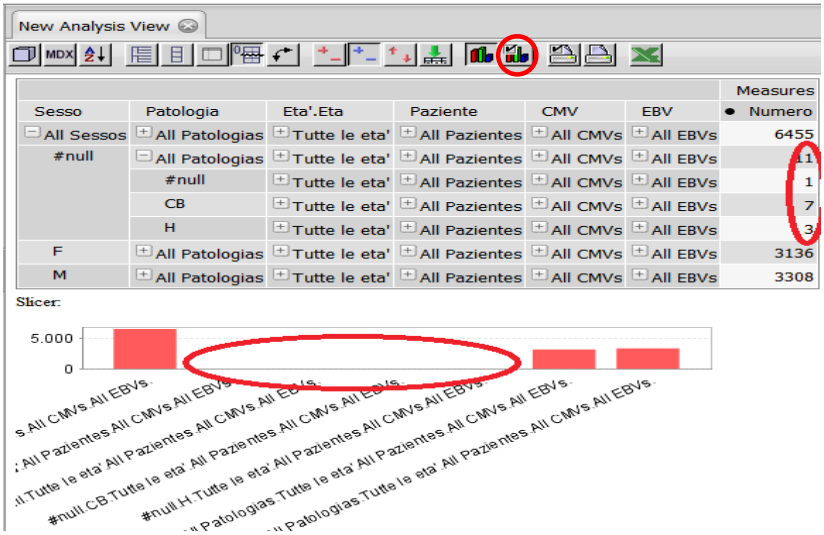


Figura 10: Dettaglio del problema della scala che si può avere nei grafici

da questa immagine è chiaro quanto sia utile avere l'interfaccia che riporta sia grafico che tabella in quanto diminuisce la possibilità di cadere in errore, in questo caso si potrebbe pensare che le barre che rappresentano i campi relativi ai "#null" siano completamente vuote mentre se si guarda la tabella si nota che ci sono pazienti di cui non si conosce il sesso che riportano diverse malattie (ciò è dovuto alle impostazioni predefinite dei grafici, che dovranno essere opportunamente adattate in base alle esigenze particolari dei grafici che vengono presi in considerazione).
Analizziamo ora a quali tipi di opzioni si può ricorrere per modificare sia il tipo di grafico sia le caratteristiche di ogni tipologia; per accedere al menu contenente le opzioni dei grafici basta cliccare sul pulsante in alto evidenziato col cerchio rosso.
Questa è la finestra che appare:

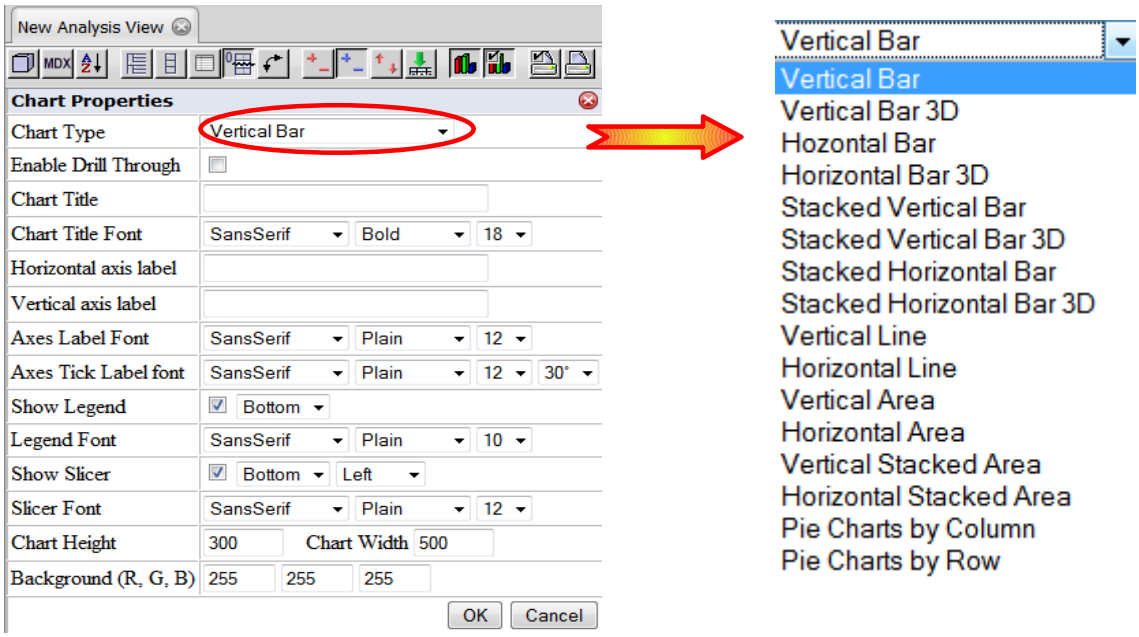


Figura 11: Lista dei grafici disponibili in Pentaho

la prima riga contiene un menù a tendina che riporta tutti i tipi di grafici che è possibile costruire con Pentaho; tra questi i più utilizzati sono gli istogrammi (Vertical Bar) e i grafici a torta (Pie Chart), notare che i grafici a torta si dividono in due categorie perchè è possibile ottenere una suddivisione della torta rispettando le righe o le colonne della tabella (più avanti questo concetto verrà spiegato più nitidamente).
Altri campi che possono tornare utili sono quelli che permettono di inserire un titolo nel grafico e di dare un nome agli assi cartesiani, è possibile inoltre modificare la legenda a proprio piacimento. Un aspetto fondamentale quando si opera con una mole ampia di dati che si vuole rappresentare graficamente è di sicuro la leggibilità del grafico, facciamo un esempio. Comprimiamo come spiegato in precedenza la parte di tabella relativa alle patologie dei "#null" e espandiamo le patologie delle pazienti femminili; facciamo un grafico a barre di questa suddivisione e quello che si ottiene è riportato nell'immagine di sinistra.

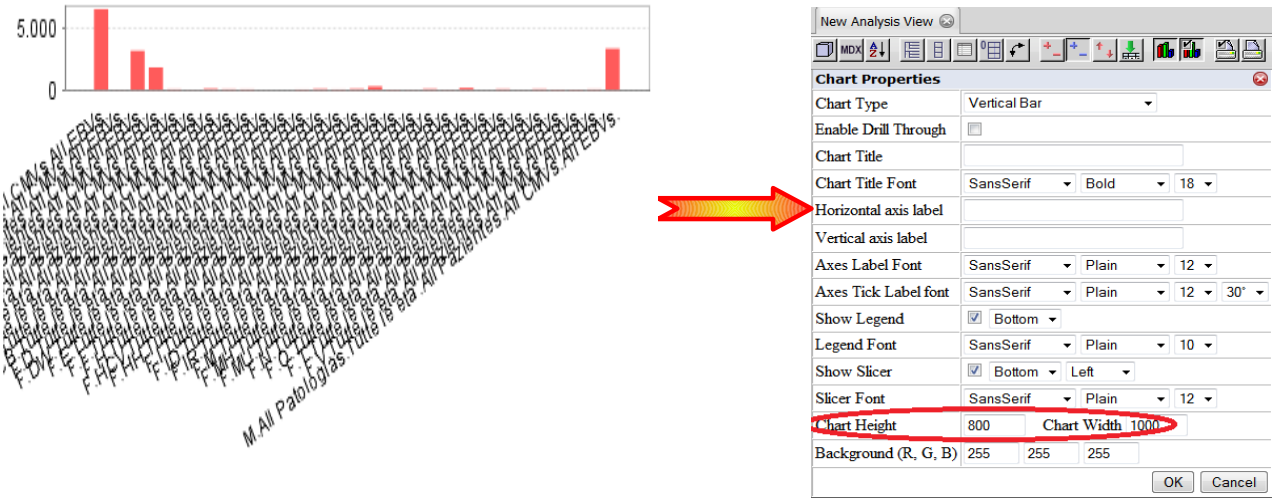


Figura 12: Modifica dei parametri che regolano altezza e larghezza del grafico

Si può notare che il grafico è praticamente illeggibile a causa della sovrapposizione delle scritte ed inoltre non si riesce a distinguere la lunghezza delle barre o a capire se ci siano o meno; quello che si può fare per risolvere il problema è modificare, nella pagina con le proprietà del grafico, la parte evidenziata in rosso che specifica l'altezza e la lunghezza del riquadro, settando come valori 800 e 1000; il risultato ottenuto è il grafico rappresentato in Figura 13.

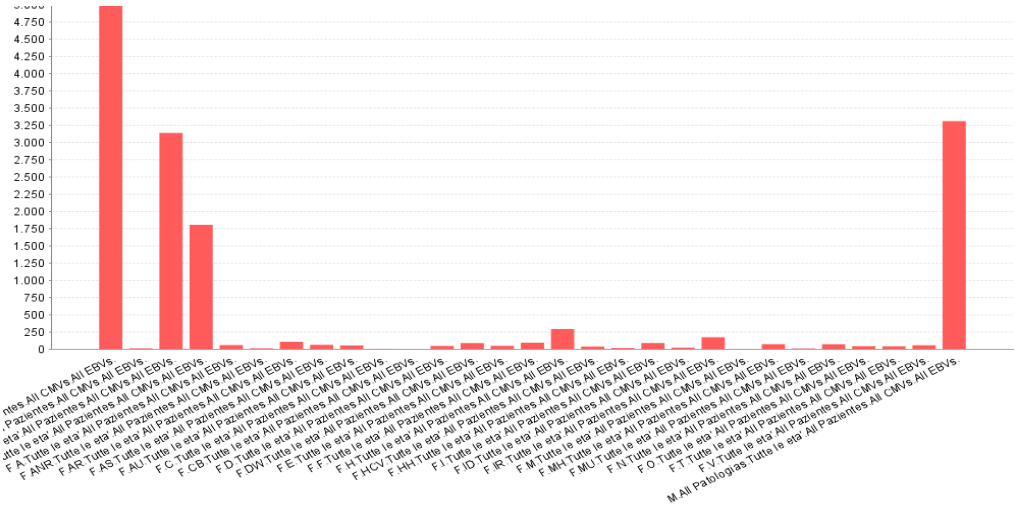


Figura 13: Grafico dopo l'ingrandimento

5. COME MODIFICARE I GRAFICI UTILIZZANDO IL “CUBO”

Prima di affrontare questa parte, sicuramente la più importante, occorre introdurre come sono strutturati i dati all'interno di questo tipo di banche dati. Il database tradizionale (E-R per intenderci) è paragonabile ad una fotografia in un determinato momento, il cubo invece raccoglie e accumula dati (come succede in un magazzino). La disposizione multidimensionale permette di confrontare più facilmente diversi dati tra loro, per generare informazioni (informazione = dato utile, sensato e riutilizzabile). Il modo più diffuso per creare questi cubi è quello che utilizza uno schema "a stella"; al centro c'è la tabella dei "fatti" che elenca i principali elementi su cui sarà costruita l'interrogazione, e collegate a questa tabella ci sono varie tabelle delle "dimensioni" che specificano come saranno aggregati i dati. Ad esempio, un archivio di clienti può essere raggruppato per città, provincia, regione; questi clienti possono essere relazionati con i prodotti ed ogni prodotto può essere raggruppato per categoria.
Per aprire l'interfaccia che permette la modifica della rappresentazione dei dati è semplice, basta cliccare sul pulsante raffigurante il “cubo” sulla sinistra dello schermo.
L'interfaccia è semplice e permette di selezionare all'interno delle varie categorie quali dati vogliono essere visualizzati in contemporanea; è possibile inoltre effettuare degli scambi tra le righe e le colonne.
Andando con ordine, per conoscere tutte le classi presenti all'interno di ogni categoria basta cliccare sulla stessa perchè si apra un menu che permette di selezionare quelle desiderate; proprio in questo modo in precedenza è stato modificato il campo Measures da “n” (neutrofili) a “Numero”.
I passi sono pochi e piuttosto semplici: aprire l'interfaccia del cubo; cliccare su “Measures”; selezionare/deselezionare i parametri d'interesse; accettare i cambiamenti cliccando su “OK” sia nel menu a tendina sia nell'interfaccia del cubo.

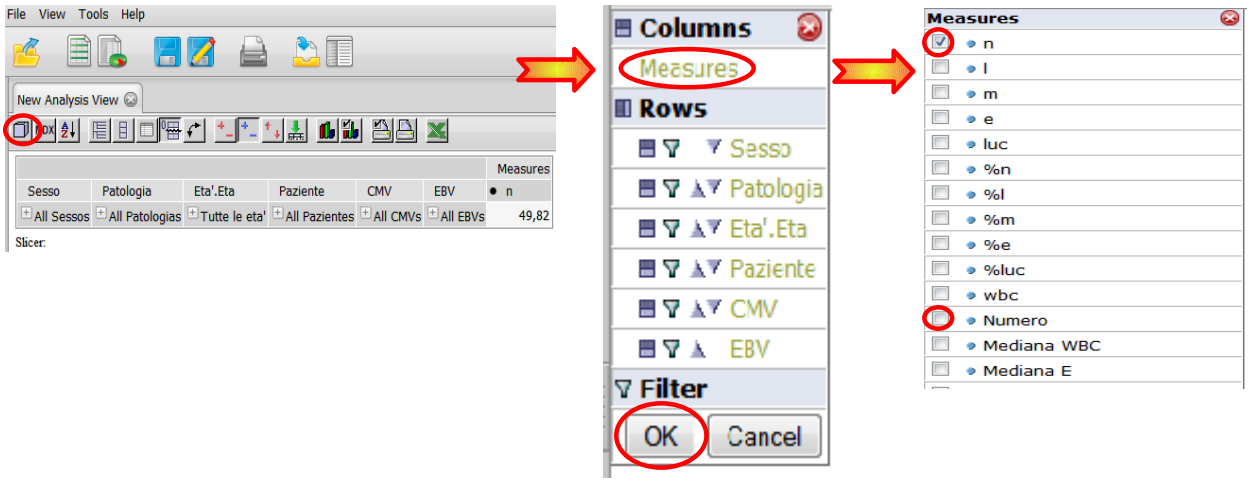


Figura 14: Sequenza per modificare i parametri da visualizzare

*(**N.B**. Da notare che quando quello che si vuole visualizzare non è il numero, ma un dato di qualche analisi in laboratorio, quello che viene fatto quando i dati sono raggruppati il valore che viene visualizzato nella tabella è una media dei singoli; inoltre questo procedimento può essere fatto su qualsiasi categoria). I possibili parametri selezionabili sono: n (neutrofili), l (linfociti), m (monociti), e (eosinofili), luc (leucociti) ed infine il numero dei pazienti.*

Un'altra cosa che può essere utile fare in alcuni casi è spostare una determinata categorizzazione da righe a colonne (o viceversa). Anche in questo caso i passi sono piuttosto semplici e intuitivi. Per prima cosa occorre aprire l'interfaccia del cubo; sulla sinistra di ogni categoria vi sono diverse icone, quella che ci interessa per ora è quella evidenziata in rosso che ci permette appunto di spostare una determinata categorizzazione, che in questo caso è applicata sulle righe, sulle colonne.

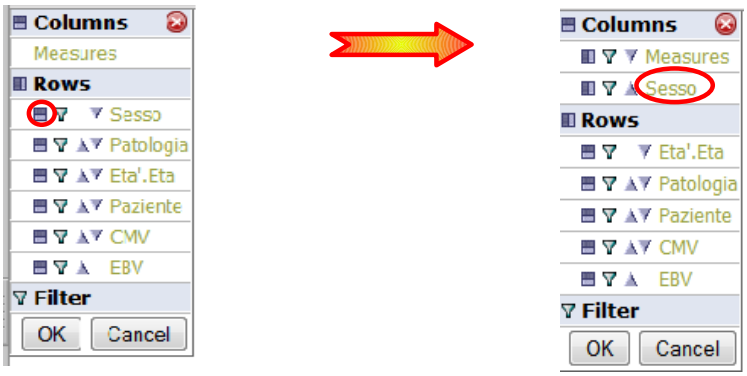


Figura 15: Spostamento categorizzazione dalle righe alle colonne

Una volta accettata la seguente modifica l'unica differenza apprezzabile visivamente sulla tabella è quella presentata in figura, ossia quella di vedere i dati separati per quella determinata categoria non sulle righe ma sulle colonne.

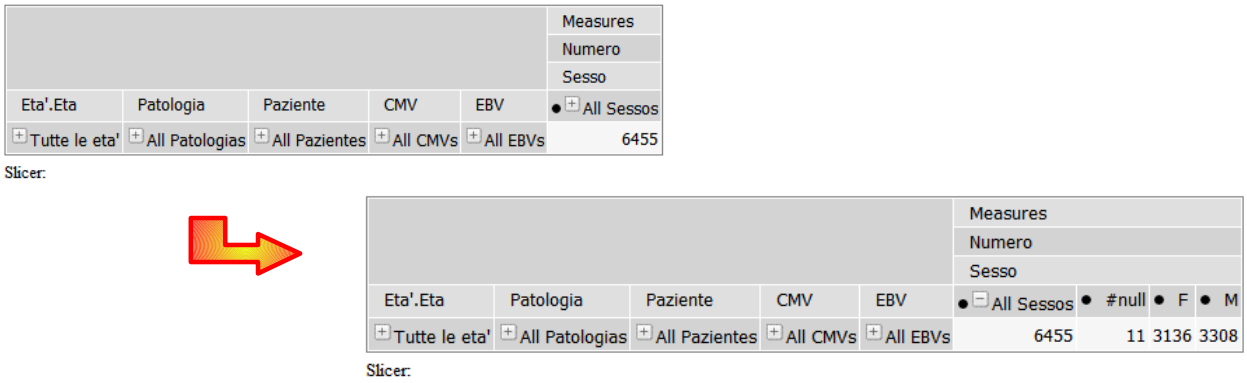


Figura 16: Come influisce lo spostamento sulla tabella

Per quel che riguarda i grafici invece si può notare un notevole cambiamento prima e dopo aver spostato righe e colonne. Nonostante l'informazione comunicata sia sempre la stessa, il modo in cui è comunicata cambia notevolmente; come si può notare dai grafici qui sotto.

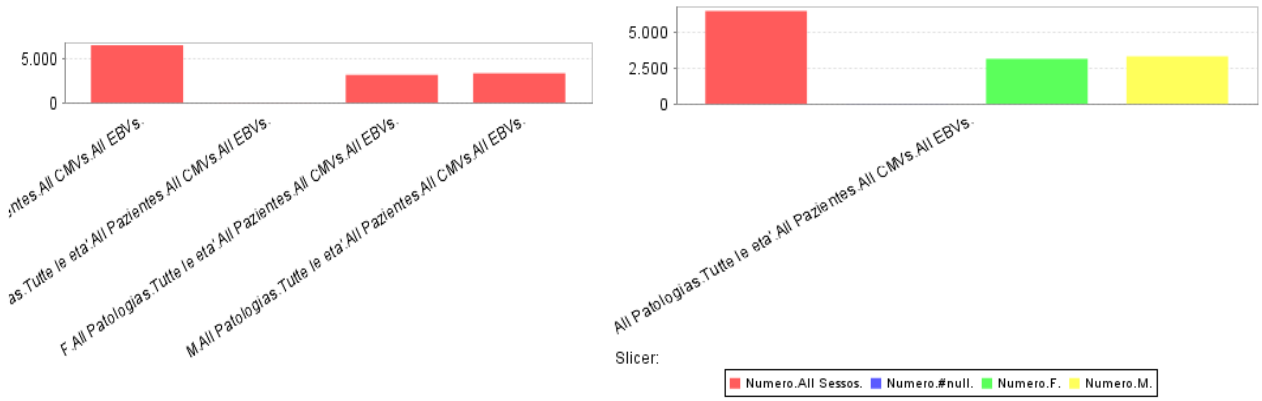


Figura 17: Esempi di grafici Prima e Dopo lo spostamento

Si può notare infatti come DOPO la categorizzazione delle colonne riporti più chiaramente i dati rappresentati nel grafico grazie a una semplice categorizzazione sulle colonne invece che sulle righe.
Una volta riportato la categoria "Sesso" tra le righe si nota come la struttura della tabella sia cambiata rispetto a quella originale, adesso infatti "Sesso" è l'ultima categorizzazione disponibile tra quelle presenti sulle righe; in Pentaho è possibile spostare le varie righe per avere in posizione agevolata le categorie che si

usano più spesso lasciando nelle ultime posizioni quelle che ricorrono meno. Per fare questa operazione si utilizza un altro dei pulsanti vicino al nome della categoria all'interno dell'interfaccia del cubo.

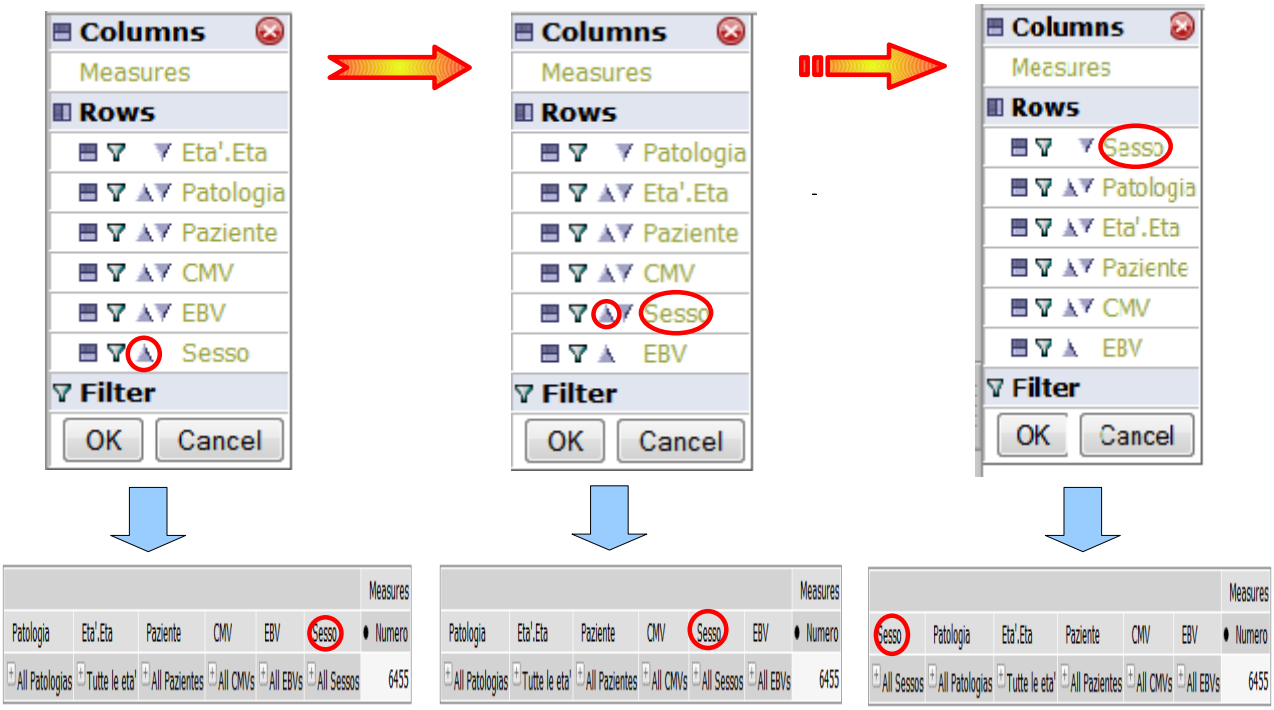


Figura 18: Sequenza di passi per spostare una categorizzazione in avanti o indietro

N.B. *Per ottenere le modifiche nella tabella occorre cliccare su "OK" ad ogni passo.*

Tutti gli strumenti utilizzati in Pentaho permettono di suddividere i pazienti tra le categorie selezionate mostrando i valori dei parametri scelti; in questo modo è possibile decidere se visualizzare ad esempio solo maschi, solo femmine, entrambi o semplicemente il totale degli individui (con "all Sessos" naturalmente). La stessa cosa può essere fatta per qualsiasi altra categoria.

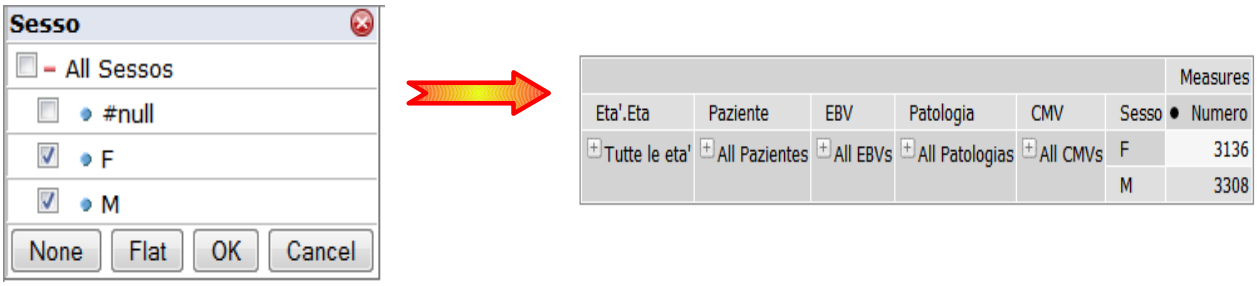


Figura 19: Effetto di una selezione sulla tabella

Il risultato di un'operazione di questo tipo è la divisione in categorie di quello che si è selezionato; quando quello che interessa non è ottenere la suddivisione in categorie ma serve sapere l'informazione complessiva della selezione desiderata si può cambiare il risultato utilizzando lo strumento "Filtro". Il filtro permette di eliminare i dati che non appartengono alla selezione d'interesse senza fare distinzioni all'interno della categoria ed eliminando la stessa dalla tabella. Come si potrà notare in seguito infatti la tabella risultante è mancante di tutte quelle categorie che vengono utilizzate per filtrare i dati.
Per utilizzare questo strumento basta cliccare sul simbolo a forma di imbuto sulla sinistra della categoria d'interesse. Per rendere l'esempio un po' più sensato lavoreremo sui CMV, i passaggi sono descritti dalle figure della prossima pagina.

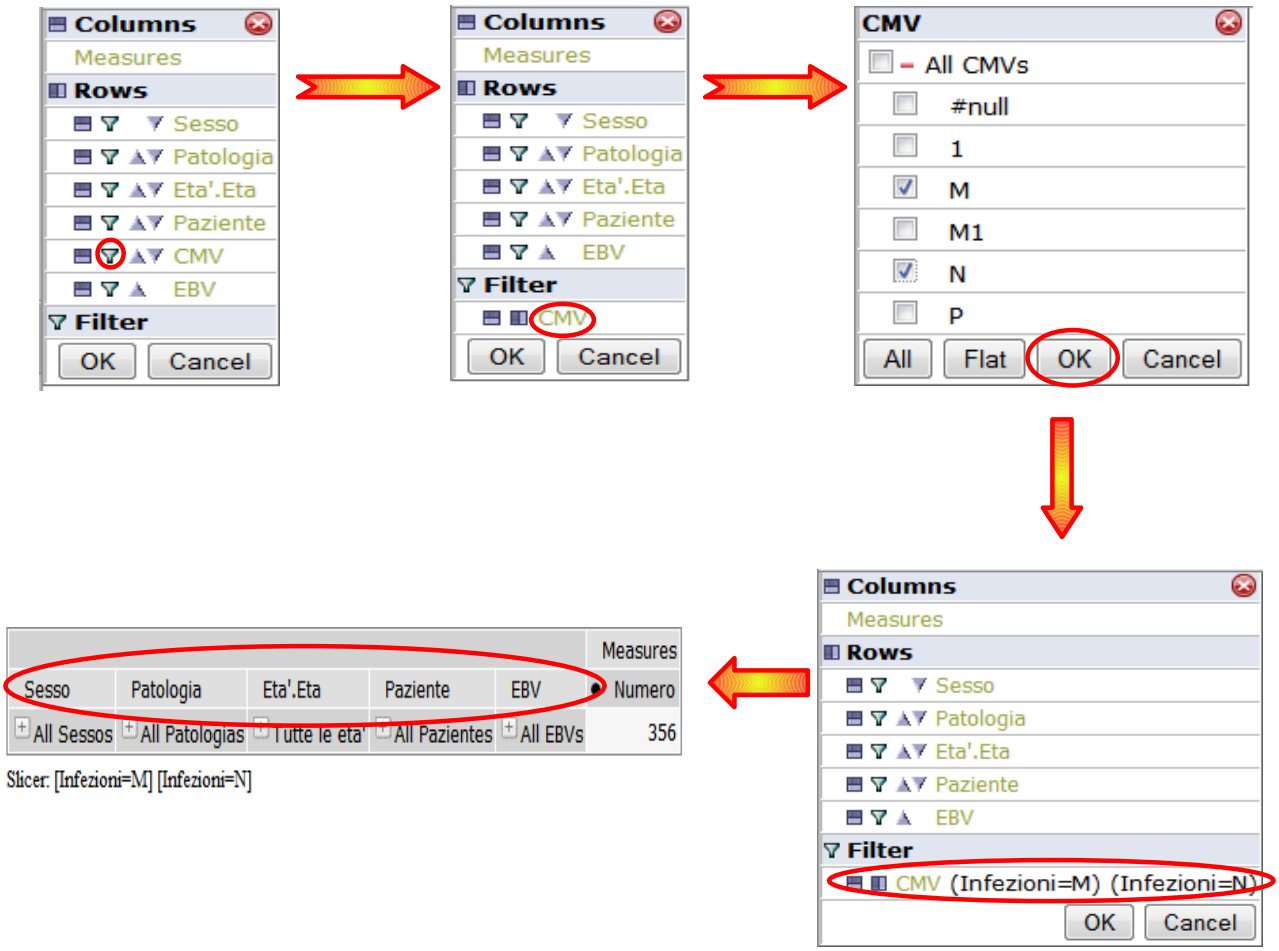


Figura 20: Applicazione di un filtro ai dati

Con questo si conclude l'elenco e la descrizione dei vari strumenti che compongono principalmente l'interfaccia di Pentaho. Seguiranno alcuni esempi che si possono creare utilizzando adeguatamente tutto ciò che è stato spiegato finora.

6. ESEMPI

Solitamente quello che può interessare è un grafico che rappresenti all'interno delle varie patologie, la suddivisione per classi di età dei pazienti e che raffiguri inoltre per ogni classe il valore di un determinato parametro. Semplifichiamo con un esempio: supponiamo che ci interessi osservare com'è distribuito il valore di un certo dato di laboratorio in una determinata sottoclasse di patologie all'interno dei pazienti di sesso maschile, supponiamo inoltre che si voglia suddividere questa distribuzione per alcune classi d'età. Quello che occorre fare è applicare tutti i concetti visti fin'ora ad un solo grafico: selezionare in “Measures” il parametro desiderato (si veda pagina 7); selezionare all'interno di “Patologia” il sottoinsieme d'interesse; spostare “Età” nelle colonne e selezionare il range che si vuole visualizzare (si vedano le pagine 7-8) ed infine filtrare i dati che verranno presentati spuntando all'interno di “Sesso” la casella che rappresenta i maschi (come spiegato precedentemente in questa pagina). Il risultato sarà simile a quello raffigurato in Figura 21.

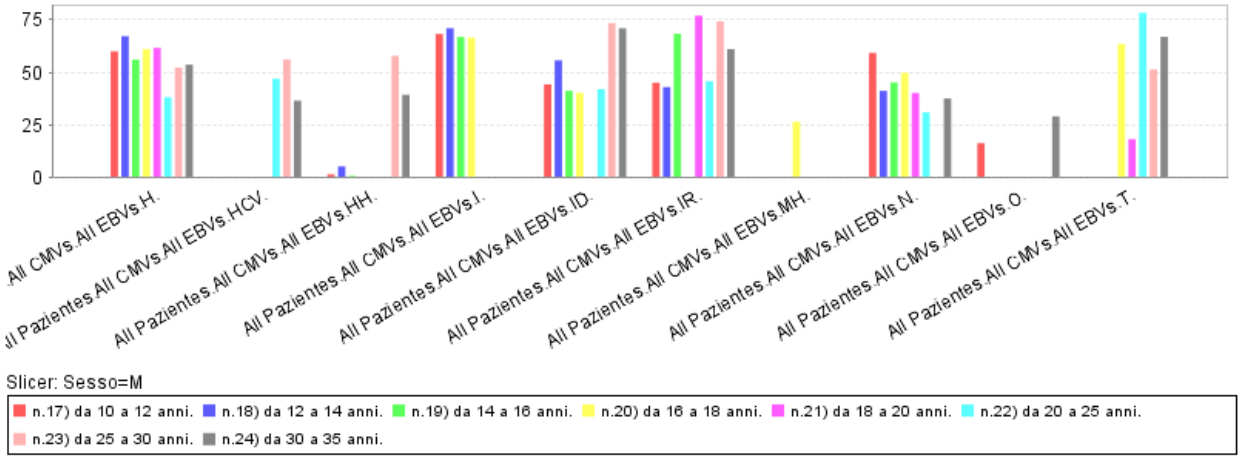


Figura 21: Grafico che rappresenta come varia il parametro n all'interno degli individui maschili affetti da una determinata classe di patologie all'interno di un certo range d'età

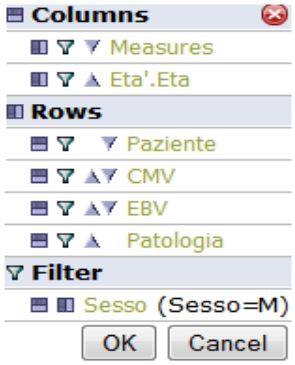


Figura 22: Cubo utilizzato per ottenere il grafico visto in precedenza

Utilizzando lo stesso scheletro del cubo (Figura 22) è possibile ottenere grafici diversi e che esprimono informazioni molto differenti. Semplicemente cambiando il parametro che si vuole visualizzare in "Numero" e selezionando un grafico a torta (Pie chart) si può ottenere il grafico in Figura 23.

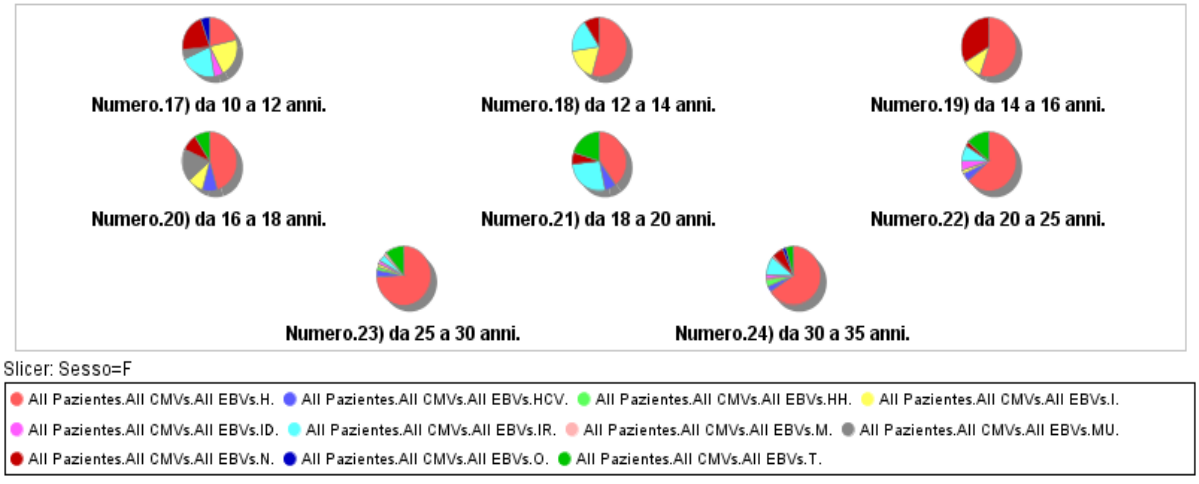


Figura 23: Grafici a torta che rappresentano il numero di pazienti di sesso femminile che presentano quella determinata malattia in quel particolare range d'età.

*(**N.B.** È consigliabile utilizzare sempre il parametro Numero con i grafici a torta in quanto rappresentano una visione d'insieme; questa visione andrebbe sicuramente persa quando si tenta di raffigurare dati di laboratorio)*

Operando nuovamente sul cubo è possibile ad esempio studiare l'andamento di una determinata patologia (o il valore di un dato parametro) col passare del tempo. Per ottenere un grafico come quello rappresentato in Figura 24 occorre prima di tutto selezionare il parametro che si intende visualizzare (solitamente "Numero"), utilizzare i filtri o selezionare all'interno delle varie categorie le sottoclassi d'interesse ed infine scegliere nell'elenco dei grafici presente in Figura 11 il grafico "Vertical Line"; questo grafico non costruisce barre bensì lega i punti contigui da una linea retta ed è molto utile se utilizzato per rappresentare dati temporali.

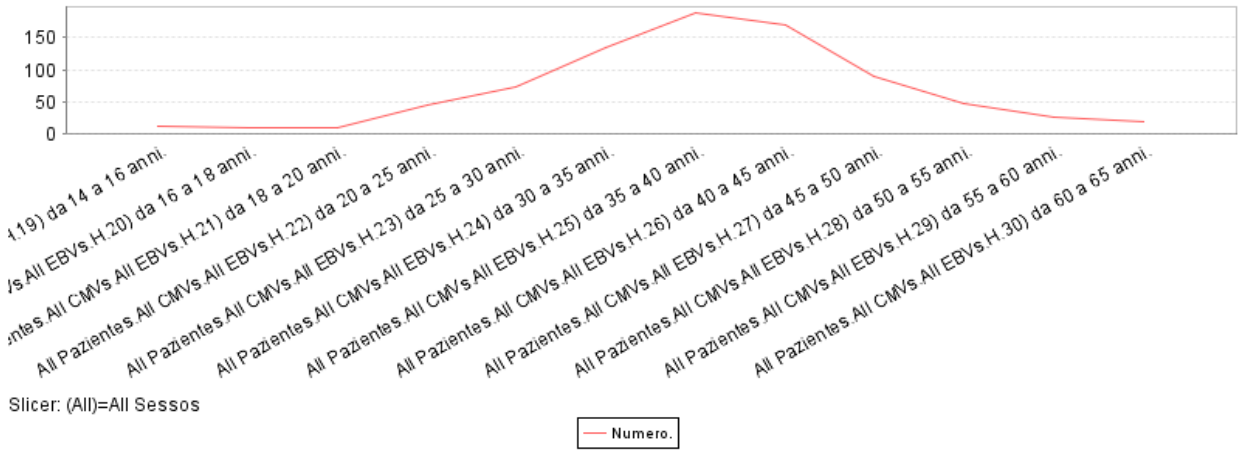


Figura 24: Grafico che rappresenta l'andamento di una determinata patologia all'interno del database secondo il passare degli anni.

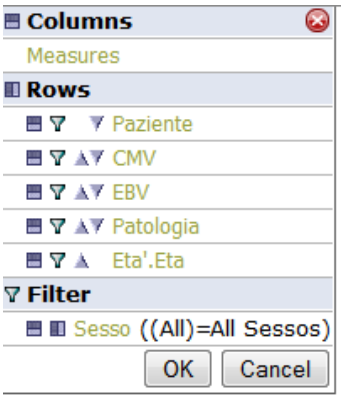


Figura 25: Visione del cubo utilizzato per ottenere il grafico della figura precedente

Con questi esempi si conclude il tutorial relativo all'utilizzo di Pentaho all'interno di analisi immunologiche.