DATASET WINE

Di seguito si riportano le osservazioni fatte durante l'analisi del dataset Wine, grazie all'impiego del software Orange Canvas.

Si è iniziato dando uno sguardo ai vari dati presenti nel database: 3 tipologie differenti di vini (la nostra variabile dipendente), e 13 attributi (le variabili indipendenti) rappresentati alcune caratteristiche chimiche del vino (intensità del colore, flavonoidi, alcol, acido malitico...).

L'obiettivo è quello di adottando l'algoritmo non supervisionato k-Means, essere in grado di predire in base alle variabili prese in esame di volta in volta, se si sta parlando del vino di tipo 1, 2 oppure 3.

Innanzitutto avendo a disposizione una cospicua quantità di covariate, si è cercato di individuare quelle che effettivamente fossero portatrici di informazione utile alla trattazione del nostro problema: non è infatti poco frequente che più dati vogliano dire in realtà la medesima cosa, risultando quindi inutili a livello pratico. Si è quindi deciso di impiegare il metodo delle *Principal Component Analysis (PCA).* Tale strategia sfrutta il fatto che i punti possono essere trasformati in vettori tramite una trasformazione lineare delle variabili; in tal modo si ottengono delle nuove variabili che vengono proiettate nei vari assi del piano cartesiano in base al risultato della loro varianza: la nuova variabile con la maggiore varianza viene proiettata sul primo asse, la seconda per dimensione della varianza, sul secondo e così via. La riduzione della complessità avviene a questo punto limitandosi a trattenere le variabili che presentano le varianze più significative.

Utilizzando il widget PCA si è provveduto a fare tale riduzione di dimensionalità: l'80% di varianza spiegata è stata ritenuta un buon valore soglia da impiegare, che ha permesso di ridurre il numero di variabili da 13 a 5.

Successivamente si è applicato l'algoritmo non supervisionato k-Means. A tal proposito si evidenzia come da un primo sguardo sulla distribuzione dei dati, si era già osservato che essi molto probabilmente risultavano essere un po' troppo "sparpagliati" per poter adattarsi completamente al suddetto algoritmo, che presenta invece l'apice della sua performance su distribuzioni a grappolo e ben concentrare. Si è comunque ritenuta questa una buona opportunità per osservare come anche in situazioni non pienamente idonee esso si sarebbe comportato, esaminando inoltre quali siano gli strumenti e le strategie che si possono impiegare per individuare i punti dove l'algoritmo di clustering fallisce.

Per avere una visione più precisa della situazione si è applicato il widget Interactive k-Means che permette di mostrare il funzionamento dell'algoritmo di clustering. È stato quindi possibile vedere come stabilendo un numero di cluster pari a 3 (equivalente al numero di tipologie di vini), questo vada a posizionare i centroidi e successivamente a distribuirvi i vari punti attorno, assegnandovi l'appartenenza ad un cluster oppure ad un altro in base alla loro distanza. Confrontando tale distribuzione con quella di partenza già a questo punto si può notare come in realtà l'algoritmo attribuisca con un certo margine di errore i punti ai vari cluster.

Successivamente si è andati ad applicare l'algoritmo supervisionato. Gli esiti sono stati esaminati singolarmente, grazie al widget SelectRow che isolando un cluster alla volta ha permesso di vedere quanti mach o missmatch si fossero verificati in rapporto col grafico originario. Ci si è inoltre fatti aiutare al widget Distributions per avere una visione ancora più chiara del tutto.

A seguire si includono alcune immagini esplicative di quanto appena enunciato.

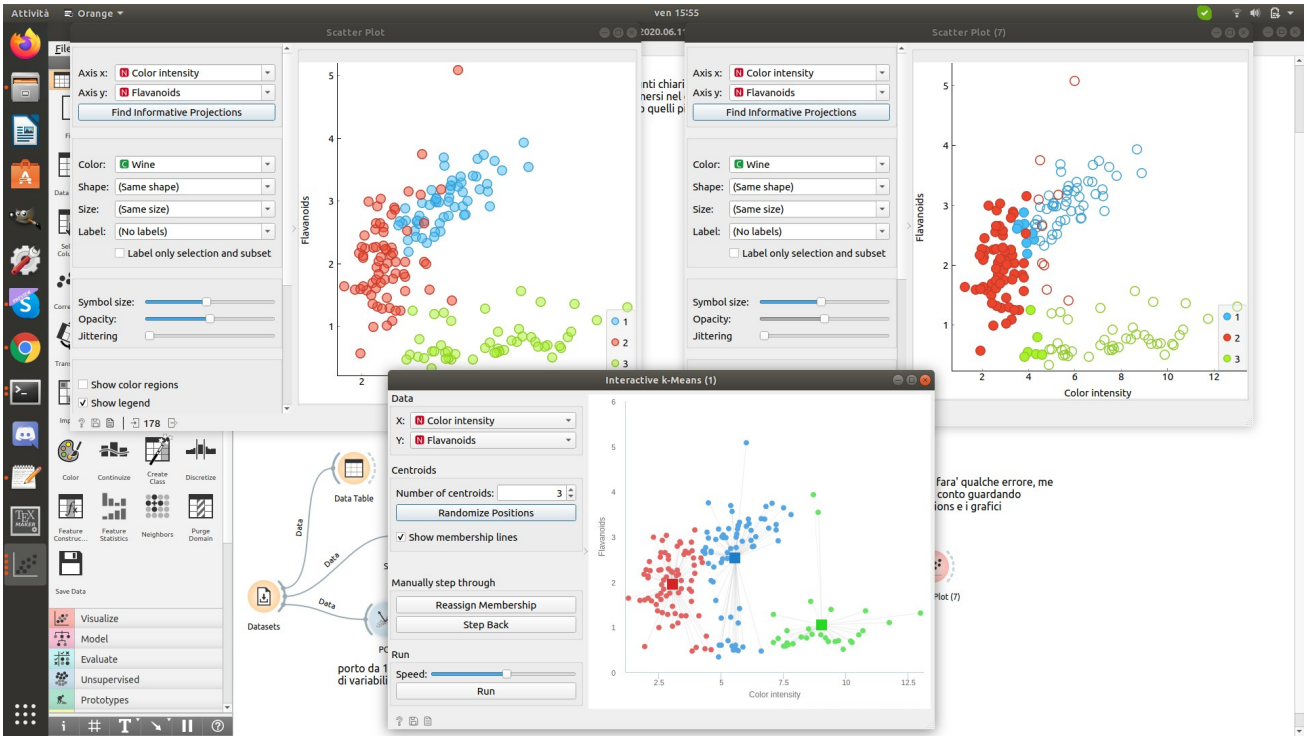

Figura 1: Grafici dei dati (Wine) – Orange Canvas

Come si può notare a ciascuna famiglia di vino corrisponde un differente colore: azzurro, rosso o verde; appare chiaro come sia complicato in realtà individuare una precisa separazione fra le classi, molti sono infatti i punti border-line o che entrano dentro l'area della classe sorella.

Nel dettaglio si presenta:

- in alto: trasformazione dei dati una volta applicata la tecnica PCA;
- sotto: redistribuzione a seguito dell'applicazione dell'algoritmo k-Means con evidenziazione dei centroidi;
- a destra: in relazione alla classe 2 (rossa), come in realtà l'algoritmo vi consideri facenti parte anche alcuni punti verdi e altri blu, escludendone altri di rossi.

Un altro argomento trattato che ha permesso di comprendere meglio il problema di attribuire ai vari punti il cluster corretto, è stato quello di Silhouette.

Indicatore di quanto un punto sia coeso rispetto al cluster attribuitogli, muovendosi in un intervallo da 0 ad 1, evidenzia come più il valore tende a crescere e più significa che l'istanza si trova circondata da elementi dello stesso cluster, più invece diminuisce più sarà evidente che si sta parlando di un valore lontano dal cluster in esame (o considerato anomalo nel caso di valori negativi).

L'immagine sotto presenta il grafico inerente alle osservazioni, prive della trasformazione PCA, e a sinistra quello di Silhouette. In particolare si sono evidenziati i valori negativi del primo tipo (azzurro), che nel grafico a destra si presentano molto vicini a quelli rossi.

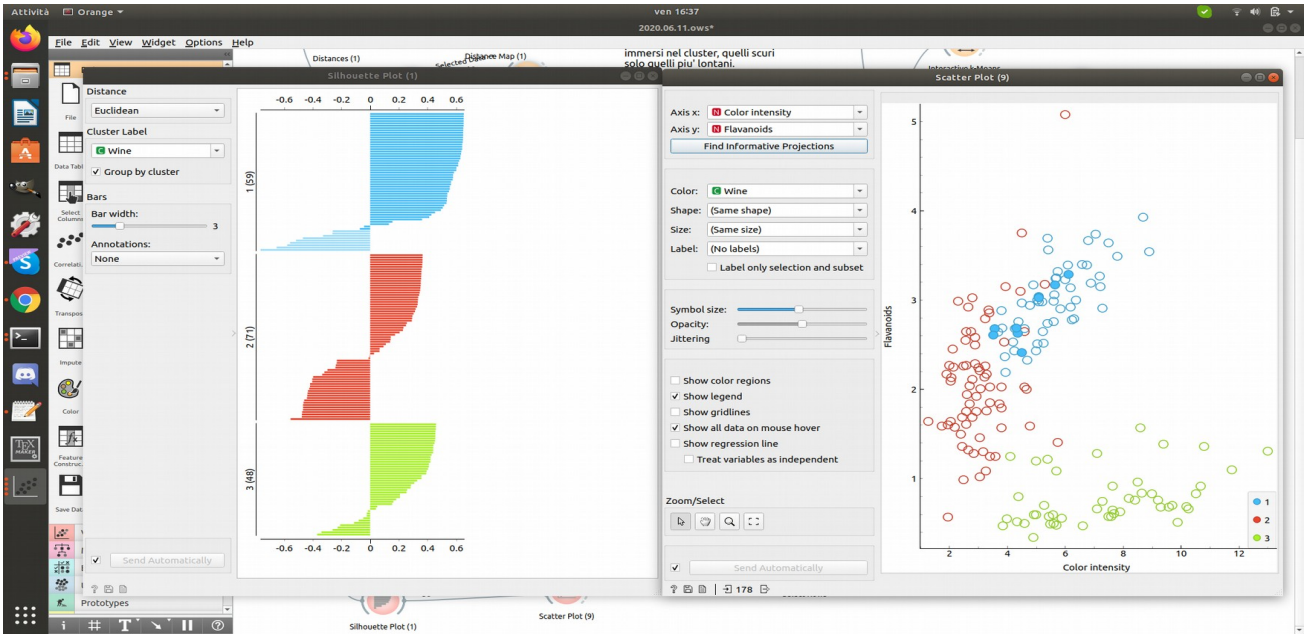

Figura 2: Silhouette Plot e relativo grafico dei dati (Wine) - Orange Canvas

In conclusione appare chiaro come l'algoritmo non supervisionato k-Means in tale circostanza faccia fatica ad individuare in modo corretto la classe di appartenenza di un dato: molte sono infatti le zone confuse che vedono la presenza di punti di colore diverso, ecco che un algoritmo basato sulle distanze non è probabilmente la scelta migliore. A seguire vengono presentati grazie al widget Distributions, le distribuzioni di ciascun cluster, dove ad esempio il cluster C3 che dovrebbe corrispondere alla classe verde, comprende anche dei punti che dovrebbero essere blu, escludendone tuttavia degli altri che dovrebbero invece farne parte.

Nelle giornate precedenti inoltre si era lavorato sul dataset Iris, simile per difficoltà di comprensione dei dati ma con una distribuzione più "a insiemi" (anche se non completamente). In tale circostanza l'algoritmo in esame aveva presentato delle prestazioni migliori.

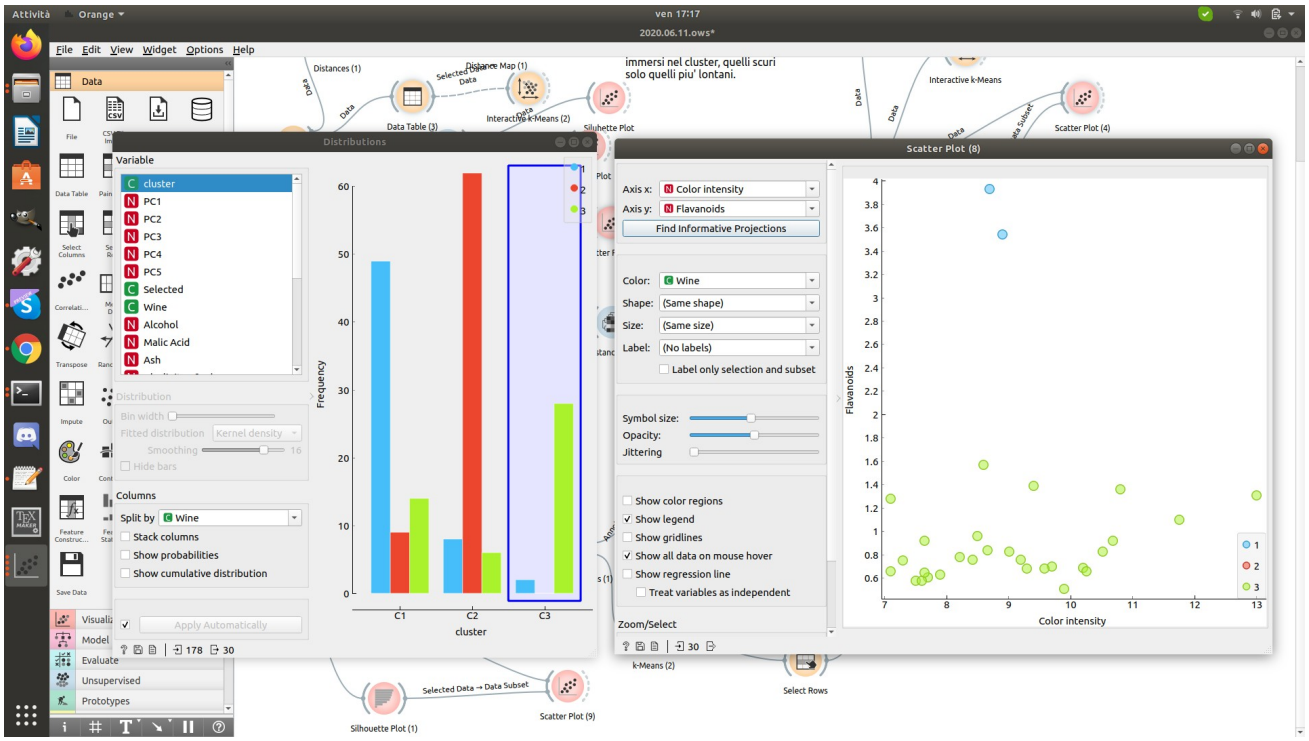

Figura 3: Dstribuzione e grafico dei dati (Wine) – Orange Canvas