**DOCUMENTAZIONE PROGETTO**

Corso di Ingegneria della Conoscenza – a.a.2020-2021

**Costruzione di una Base di Conoscenza ed utilizzo di tecniche di apprendimento supervisionato e non per l’analisi del gradimento delle *Airbnb rentals***

**Studenti:**

Barbaro Nicola – mat. 665975

Gasbarro Roberto – mat. 652507

1. **La *mission***

L’obiettivo di questo progetto è stato quello di analizzare un dataset di camere (*rentals*) proposte da Airbnb, un marketplace per il lodging, per poter offrire all’utente la possibilità di effettuare la ricerca di camere che più rispecchiano le sue esigenze: l’utente infatti è abilitato ad effettuare query personalizzate. Il secondo obiettivo è stato quello di trovare gruppi di camere simili tra loro attraverso tecniche di apprendimento non supervisionato. Infine, è stata addestrata una Belief Network per capire le relazioni che sussistono tra le varie features della camera e l’indice di gradimento di quest’ultima, per permettere all’utente di predire la probabilità di gradimento di una camera in relazione alle preferenze da lui indicate.

1. **Linguaggi e strumenti utilizzati**

Per la realizzazione del progetto è stato utilizzato principalmente il linguaggio Python, in quanto dispone di molte librerie che consentono di lavorare agevolmente con i dati, di utilizzare tecniche di apprendimento automatico e di costruire e utilizzare reti bayesiane. In particolare, sono state usate le seguenti librerie:

* Scikit-learn e Scikit-learn-extra, per l’utilizzo di modelli di apprendimento automatico e per la riduzione delle features;
* Pandas, per l’utilizzo dei Dataframe, che permettono agilmente di gestire grandi moli di dati;
* Pyswip, per l’interazione con il programma SWI-Prolog (richiede la presenza del software SWI-Prolog sulla propria macchina: <https://www.swi-prolog.org/Download.html>);
* Pgmpy, per la costruzione e manipolazione di Belief Network e per l’inferenza probabilistica.

È stato utilizzato, inoltre, il linguaggio Prolog, che adotta il paradigma della programmazione logica, per poter costruire una base di conoscenza a partire dal dataset e per poter sottoporre delle query all’agente.

Infine, l’utilizzo del software Weka ha permesso di addestrare una rete Bayesiana a partire dal dataset: di questa vengono apprese non solo la struttura ma anche le probabilità.

1. **Fase di preprocessing**

Il dataset è stato estratto dal sito *Insideairbnb*. Il dataset è formato da 74 features (contenenti informazioni dettagliate su ogni camera) e da 29581 osservazioni, aggiornate al 4 Agosto 2021.

La fase di preprocessing si articola in tre parti:

* 1. **Data Cleaning**  
     questa fase si concentra inizialmente sull’eliminazione di colonne non importanti ai fini dello studio; la lista contiene, ma non si limita, posizione geografica delle camere, numero massimo di pernottamenti, informazioni superflue sul profilo dell’host. Segue una selezione dei tipi principali di proprietà e dei quartieri maggiormente presenti all’interno del dataset.  
     Alcuni valori vuoti sono stati sostituiti con la mediana degli altri valori nel caso di variabili di tipo numerico ovvero con la moda nel caso di variabili di tipo categorico.  
     In questa fase viene generato il file pre-processato per la costruzione della Base di Conoscenza in Prolog, che verrà completato successivamente con l’aggiunta della colonna *n\_cluster*. Dopo aver estratto i dummies delle variabili categoriche, si passa alla fase successiva
  2. **Creazione della matrice Termine-Documento per le amenities**L’obiettivo è creare due TD-Matrix: la prima contenente tutte le possibili amenities estraibili dal dataset (1028) e destinata al calcolo dei clusters; la seconda contenente un numero ridotto di amenities (solo coloro che appaiono un numero di volte almeno pari alla radice quadrata del numero di osservazioni) e destinata alla costruzione della rete Bayesiana.  
     Il dataset presenta le amenities all’interno di un’unica feature, nella forma di una stringa di testo. In primo luogo, l’agente si assicura dell’assenza di amenities ripetitive o semanticamente simili. Segue poi la creazione della matrice sparsa riportante l’indice della camera sulle righe e le amenities sulle colonne.  
     Viene eliminata la colonna *amenities* e sostituita con la matrice termine-documento.
  3. **Raffinamento dei dati**Vengono discretizzate in classi le features *price, minimum\_nights* e *number\_of\_reviews* nella copia del dataset destinata alla creazione della rete Bayesiana. Vengono poi studiati, nel dataset principale, gli outliers, eliminando tutte le features presentanti un numero di outliers maggiore del 20%.  
     Segue la riduzione del numero di features del dataset del calcolo dei cluster a 26, utilizzando tecniche di Principal Component Analysis. Il numero di features da mantenere è stato scelto in base alla riduzione della varianza, come riportato nel grafico.  
     In seguito è stata effettuata la normalizzazione del dataset destinato al calcolo dei cluster attraverso l’uso del MinMaxScaler.

**MinMaxScaler**

L’algoritmo MinMaxScaler prende in input ogni valore di una feature, ne sottrae il valore minimo e lo divide per il range, ovvero la distanza tra il valore massimo e minimo originari. MinMaxScaler preserva la curva della distribuzione originaria, senza perdere le informazioni semantiche dei dati originari.

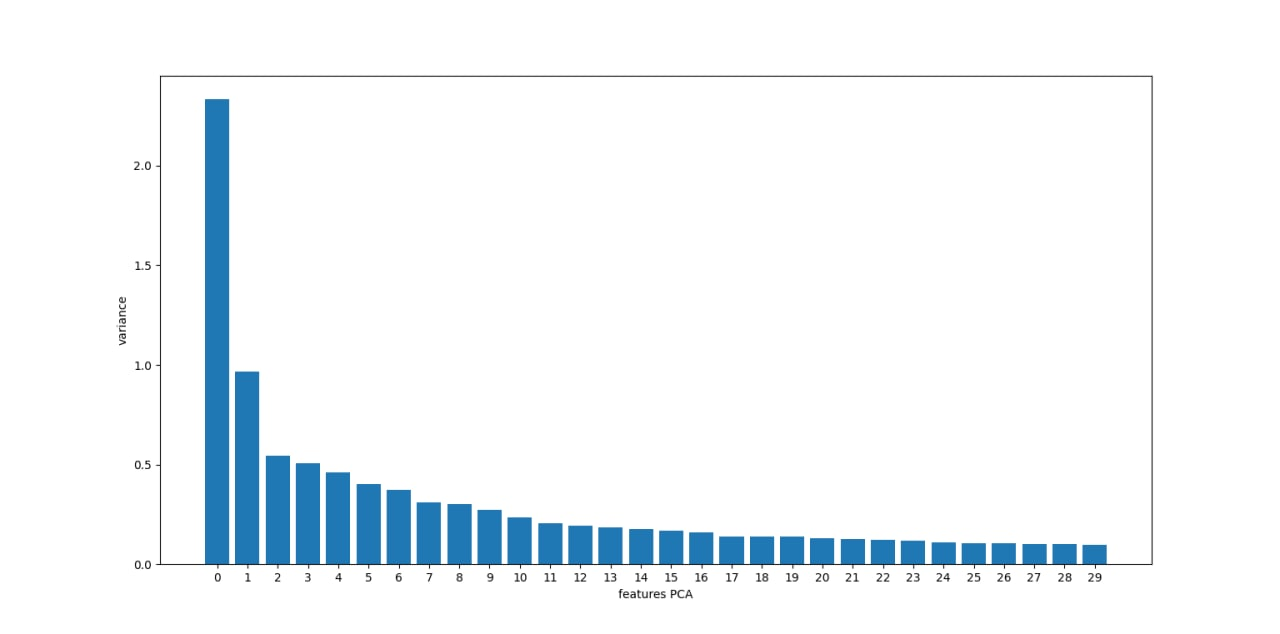


Figura 2-Scelta del numero di Features per PCA

1. **Clustering**

Si è voluto sfruttare una tecnica di apprendimento non supervisionato per individuare indicativamente le classi di camere più simili tra loro, per agevolare lo studio dei pattern di features che influenzassero di più il rating della camera.

La tecnica scelta è stata quella del Clustering.

Per stabilire quale sia il k migliore con cui effettuare il clustering, si utilizza il metodo dell’elbow: si esegue un plot dell’errore ottenuto un funzione di k e il k migliore è quello per cui successivamente l’errore diminuisce in maniera non rilevante. In corrispondenza del k migliore, quindi, sul grafico si viene a creare un vero e proprio “gomito”.

Inizialmente è stata adottata la tecnica del K-Means: essa mira a partizionare le osservazioni in un numero predefinito di clusters, creando dei centroidi, ovvero dei prototipi del cluster stesso, e assegnando il valore di cluster ad ogni esempio in base al centroide più vicino in termini del quadrato della distanza Euclidea.

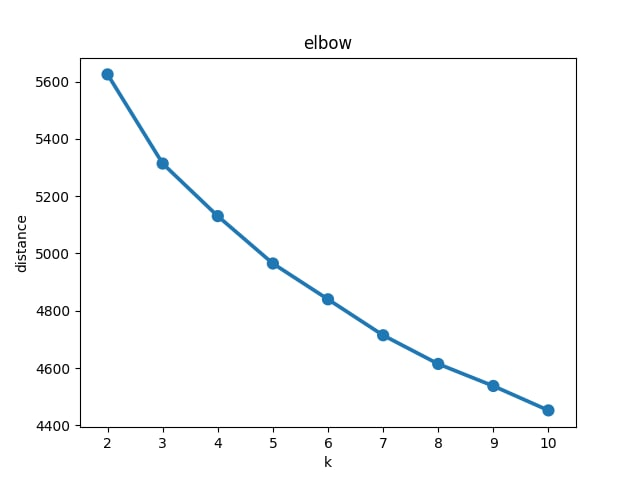
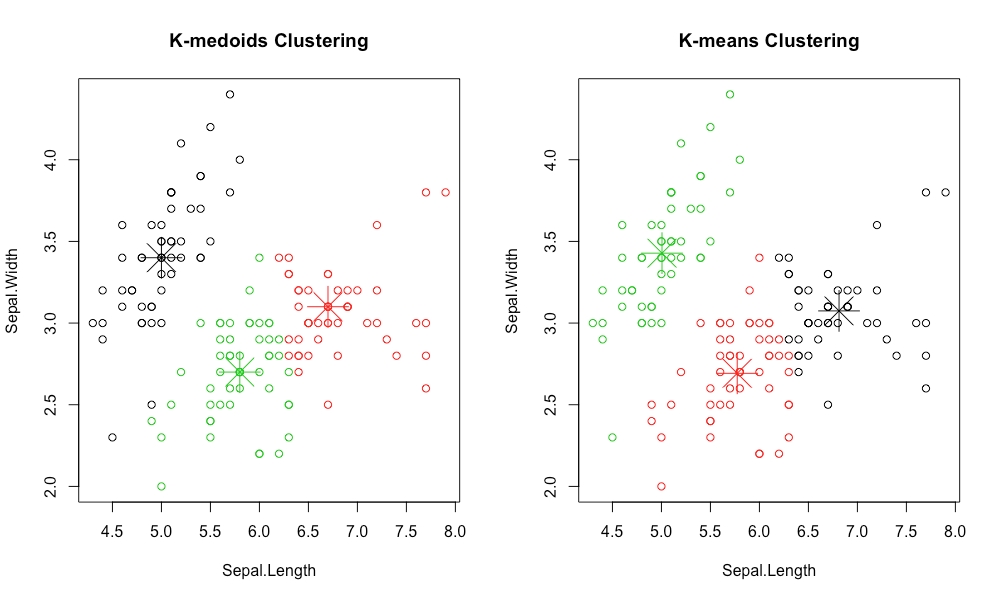
Si riporta il grafico ottenuto con K-Means, facendo variare il numero di cluster k da 2 a 10. 

Figura - K-Means

Come si evince dal grafico, non si apprezza una sufficiente tendenza del dataset a clusterizzare. Inoltre, il K-Means presenta una difficoltà a livello semantico: i centroidi trovati rappresentano la media dei valori delle features degli esempi appartenenti a quel determinato cluster, per cui il centroide potrebbe non corrispondere a nessuno degli esempi del dataset. Per questo motivo, si è scelto di virare verso una tecnica alternativa di clustering, chiamata K-Medoids.

Un *Medoid* di un cluster è definito come l’esempio all’interno del cluster stesso la cui dissimilarità media rispetto a tutti gli altri esempi del cluster è minimale: esso è il punto più centrale del cluster. La differenza sostanziale con il K-Means, oltre a presentare un algoritmo di calcolo leggermente modificato e una complessità leggermente maggiore, è che un *Medoid* è un esempio effettivo del dataset.



Si riporta il grafico ottenuto con K-Medoids, facendo variare il numero di cluster k da 2 a 10.

