**Sorrento Marina**

Progetto di Fondamenti di Intelligenza Artificiale

Versione 0.1

|  |  |
| --- | --- |
| Riferimento | FIA |
| Versione | 0.1 |
| Data | 05/02/2021 |
| Destinatario | Prof. Fabio Palomba, Prof. Fabio Narducci |
| Presentato da | Tutto il team |

# Composizione Team

|  |  |  |
| --- | --- | --- |
| Nome e Cognome | Matricola | Acronimo |
| Luigi Maiorano | 0512105861 | LM |
| Daniele Carangelo | 0512105819 | DC |
| Francesco Pio Covino | 0512105735 | FPC |

Sommario

[Composizione Team 2](#_Toc63502294)

[Revision History 3](#_Toc63502295)

[1 Introduzione 4](#_Toc63502296)

[2 Definizione dell’ambiente 4](#_Toc63502297)

[3 Raccolta dei Dati 4](#_Toc63502298)

[4 Tipologia di apprendimento 4](#_Toc63502299)

[5 Implementazione 4](#_Toc63502300)

[6 Valutazioni finali e performance 4](#_Toc63502301)

# Introduzione

Il sistema “SorrentoMarina” è una piattaforma web che si occupa della gestione di un consorzio di lidi, il cui scopo principale è permettere una gestione efficace ed automatizzata delle prenotazioni in modo che il turista possa in pochi e semplici click riservarsi un posto nel lido da lui richiesto. In questo ambito si colloca il nostro modulo di Intelligenza Artificiale che aiuta e guida il turista nella scelta del lido che più rispecchia le sue esigenze in termini di servizi disponibili, disponibilità, costi, attività proposte ….

Il modulo di profilazione proposto farà uso di tecniche ed algoritmi di Machine Learning per costruire un sistema di suggerimenti che sia il più affidabile possibile. La definizione dell’agente intelligente dovrà far capo a delle scelte cruciali:

in quale ambiente dovrà agire, quali caratteristiche dei dati deve prendere in considerazione, a quale tipo di conoscenza può fare riferimento e come il suo operato viene valutato.

Saranno tutti problemi e obiettivi che il nostro team si è imposto di analizzare e implementare.

# Definizione dell’ambiente

Ovviamente la prima fase della nostra analisi non poteva non essere lo studio e la descrizione dell’ambiente e come il nostro modulo si collocasse in esso. Per fare ciò abbiamo prima di tutto usato il modello PEAS (Performance, Environment, Actuators, Sensors) per schematizzare quelle che sono le sue caratteristiche principali:

* **Prestazioni:** Le prestazioni saranno misurate in funzione dell’accuratezza con cui l’agente riuscirà a consigliare al turista il lido da lui più adatto e quindi del gradimento globale della sua esperienza presso tale struttura, caratteristica osservabile anche da un eventuale recensione effettuata a posteriori. Per approfondire tale aspetto si rimanda alla sezione 6 in cui sono esposte valutazioni e misure di performance finali.
* **Ambiente:** Gli elementi fondamentali con cui l’agente opera sono ovviamente i turisti e i lidi.
* **Attuatori:** Lo strumento che permette all’agente di interfacciarsi con il turista è una sezione apposita denominata “Lidi Consigliati” che permetta proprio di mettere in evidenza quelli che sono i lidi consigliati dall’agente.
* **Sensori:** Lo strumento che permette invece all’agente di ottenere informazioni sul turista e le sue preferenze è una sezione apposita che rimanda ad un questionario, con apposite domande per ricavare il tutto e dare il via all’operato dell’agente.

# Raccolta dei Dati

Durante questa fase di raccolta ci siamo concentrati sul ricavare e analizzare i dati provenienti da potenziali turisti e più nello specifico comprendere quali fossero le caratteristiche fondamentali nella valutazione di un lido.

Siamo partiti dal concetto di profilo: identificare di quali aspetti di un turista tener conto per iniziare a creare un Dataset di partenza.

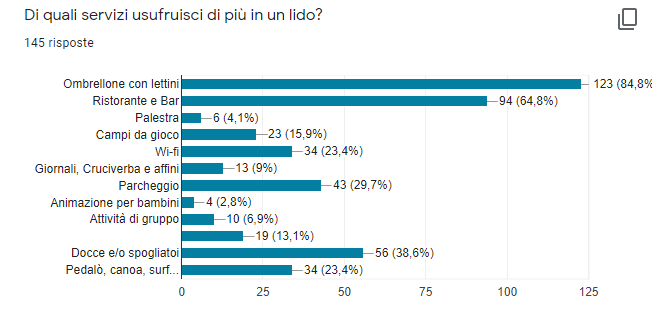
Per favorire la raccolta di informazioni abbiamo creato un questionario, disponibile al seguente [link](https://forms.gle/ABqsNTrP9Bn3Z46v9).

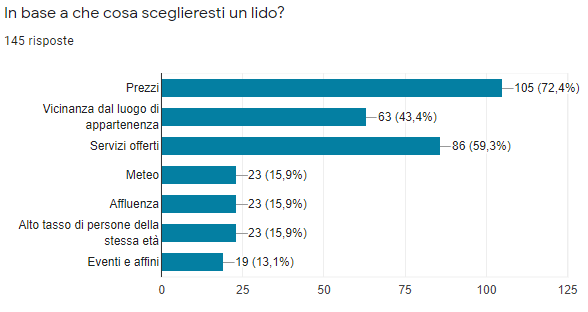
Le domande scelte erano sui seguenti argomenti:

* L’età
* Il sesso
* La nazionalità
* La situazione sentimentale
* La presenza o meno di figli minorenni
* Il lavoro attuale
* Componenti della famiglia con disabilità
* La presenza di animali domestici
* Hobby (sport, film, ballo, cucina, e cosi discorrendo...)
* Tempo di soggiorno in un lido
* Con chi si trascorrono le vacanze al mare
* Mezzo per raggiungere la spiaggia
* Servizi più usati in spiaggia
* Caratteristiche che un lido dovrebbe avere

La raccolta dei dati ha dato esito positivo, le risposte erano ben distribuite per le varie opzioni, abbiamo deciso di fermare la raccolta a quota 145 campioni.

Due esempi dei risultati ottenuti:





Da come si evince sopra sono risultati vincenti, nella scelta del lido, caratteristiche quali il prezzo del lido e i servizi che quest’ultimo può offrire al cliente.

Considerando il dominio in cui l’agente opera, un ambiente balneare ristretto, abbiamo scelto, una volta finita la raccolta dei dati, di fare 2 accorgimenti:

1. Non considerare la nazionalità, in quanto inutile ai fini del sistema.
2. Trasformare gli attributi multi valore, quali hobby o servizi preferiti, in singoli attributi binari. Questa trasformazione ha il fine di rendere più semplice la visualizzazione dei dati.

Alla conclusione dell’analisi siamo arrivati a considerare 145 istanze, ognuna su 21 attributi come mostrato di seguito:





# Tipologia di apprendimento

A causa della completa mancanza di conoscenza a priori, in quanto non è possibile in alcun modo conoscere quale lido appartiene ad una classe apprezzata dal turista, abbiamo deciso di implementare un algoritmo di apprendimento non supervisionato.

Abbiamo deciso di affidarci, data la sua semplicità, al K-Means per effettuare un clustering sui dati.

Il K-Means ha come obiettivo creare cluster (gruppi) di campioni che abbiano caratteristiche simili partendo da un dataset iniziale di addestramento, non etichettato.

Il K-Means ha una proprietà fondamentale: il numero di cluster è definito in anticipo.

Abbiamo deciso di implementare sia il K-Means che il DBSCAN, che effettua il clustering sul concetto di densità. La scelta è stata fatta confrontando la qualità dei cluster e il coefficiente di forma degli stessi.

# Implementazione

Il codice in .py è presente al seguente link [Github](https://github.com/GPS-IS-The-Boys/FIA).

Il primo step implementativo è stato quello di pulire il dataset ottenuto dal modulo. Abbiamo scelto e pulito i dati da tuple contenenti valori nulli (NaN) che rendevano difficile il trattamento degli stessi .

Lo step successivo è stato quello di eliminare le colonne testuali che influivano negativamente sul trattamento dei dati, quindi abbiamo eliminato le colonne 'Anni','Sesso','SituazioneSentimentale','Figli', 'Lavoro'.

Infine, il dataset utilizzato è stato adattato per poter essere processato e ciò è stato fatto rimuovendo le caratteristiche non discrete come l’id.

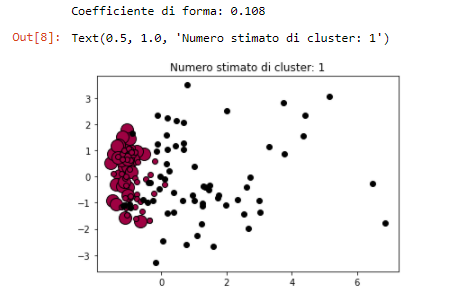
Durante la fase di clustering abbiamo deciso di implementare entrambi gli algoritmi K-MEANS e DBSCAN, implementati utilizzando la libreria sklearn.cluster.

Per quanto riguarda il DBSCAN i dati sono stati inizialmente scalati e la distribuzione normalizzata.

Successivamente a causa della presenza di più caratteristiche discrete abbiamo utilizzato l’analisi delle componenti principali (PCA) per poter estrarre tutte le caratteristiche non concorrenti e quindi riducendo la complessità del problema passando da uno spazio di rappresentazione più consono.

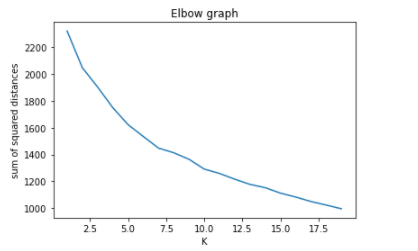
I dati, finita la fase di scaling con PCA, sono stati dati in input alla funzione di fitness del DBSCAN, la quale calcola la distanza di ogni dato dall’altro secondo la distanza euclidea.

Nella seguente immagine il risultato ottenuto dall’applicazione del DBSCAN

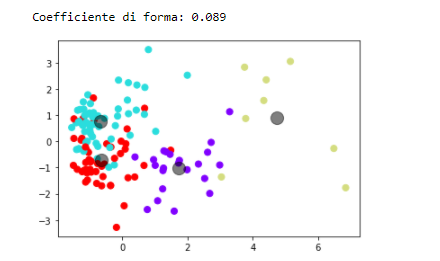


Per quanto riguarda l’algoritmo K-MEANS non c’è stato bisogno di nessun tipo di pre-processing ulteriore. Anche in questo caso la funzione di fitness calcola una distanza euclidea.

La scelta di K è avvenuta con il metodo del gomito tramite il calcolo della somma dei quadrati degli errori, di cui sotto il risultato.



Da cui abbiamo deciso di applicare il K-Means prima con k = 4 e poi con k = 5 per scegliere, infine, la prima opzione, come di seguito illustrato (i punti neri rappresentano i centroidi).



# Considerazioni finali

Una volta completata la fase di clustering in cui sostanzialmente abbiamo riunito gli elementi che sono più simili rispetto alle risposte date non ci resta che comprendere quale lido è più adatto ai gusti degli utenti.

Il Main è organizzato nel seguente modo:

1. Il sistema pone all’utente delle domande per memorizzare le risposte (le stesse poste nel questionario)
2. L'utente risponde
3. Le risposte vengono usate come dati di testing e date in pasto all’algoritmo
4. L'algoritmo ritorna i lidi che secondo il clustering possono essere associati ai gusti e alle caratteristiche dell’utente

Per quanto riguarda le performance del sistema proposto ci sono due fattori che abbiamo tenuto in considerazione e che abbiamo cercato di massimizzare:

1. Il coefficiente di forma di ogni cluster formato durante la fase di clustering
2. L’accuratezza dei suggerimenti proposti

Il primo parametro deriva dall’analisi del coefficiente di forma che sostanzialmente indica il livello di coesione e di separazione dei dati all’interno di ogni cluster.

I risultati raccolti durante la fase di apprendimento ci hanno indirizzato nella scelta dell’algoritmo tra i due scelti, da utilizzare.

Il Clustering effettuato dall’algoritmo DBSCAN ha mostrato un coefficiente di forma 0,203 fatto da 1 cluster e con 39 punti di rumore( eps = 4 e min\_samples = 12).

Il K-means ha invece ottenuto un coefficiente di forma di 0,088, dimostrandosi decisamente migliori. La possibilità di scegliere a priori il numero di cluster ha favorito la scelta, dato che il DBSCAN ne aveva previsto uno solo.