ANALISI 'DRINK DATASET'

Simone Capucci

DATASET INIZIALE

Il dataset è composto da **546 variabili** (bevande) con **340 attributi**: Drink, IdDrink, Alcoholic, Category Glass e i vari ingredienti (rappresentati in colonne binomiali). Successivamente ho reso binomiali anche le colonne Alcoholic, Category e Glass

```
Number of instances = 546
Number of attributes = 340
Dimensions of dataset 546 x 340
```

	Drink	idDrink	Alcoholic	Category	Glass	Peppermint extract	Coconut syrup	Jello	Powdered sugar	caramel sauce	...	Nutmeg	Lime peel	Food coloring	Light cream	Applejack	Grand Marnier	Cardamom	Marshmallows
0	'57 chevy with a white license plate	14029	alcoholic	cocktail	highball glass	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	1-900-fuk-meup	15395	alcoholic	shot	old-fashioned glass	0	0	0	0	0	...	0	0	0	0	0	1	0	0
2	110 in the shade	15423	alcoholic	beer	beer glass	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	151 florida bushwacker	14588	alcoholic	milk / float / shake	beer mug	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	155 belmont	15346	alcoholic	cocktail	white wine glass	0	0	0	0	0	...	0	0	0	0	0	0	0	0

5 rows × 340 columns

DATA EXPLORATION

```
alcoholic=non alcoholic
0    488
1     58
Name: count, dtype: int64
```

ABBIAMO %MAGGIORE DI BEVANDE ALCOLICHE

```
glass=highball glass
0    440
1    106
Name: count, dtype: int64
```

CI SONO MOLTE BEVANDE CON HIGHBALL GLASS

```
category=coffee / tea
0    521
1     25
Name: count, dtype: int64
```

25 TIPOLOGIE DI CAFFE' - THE

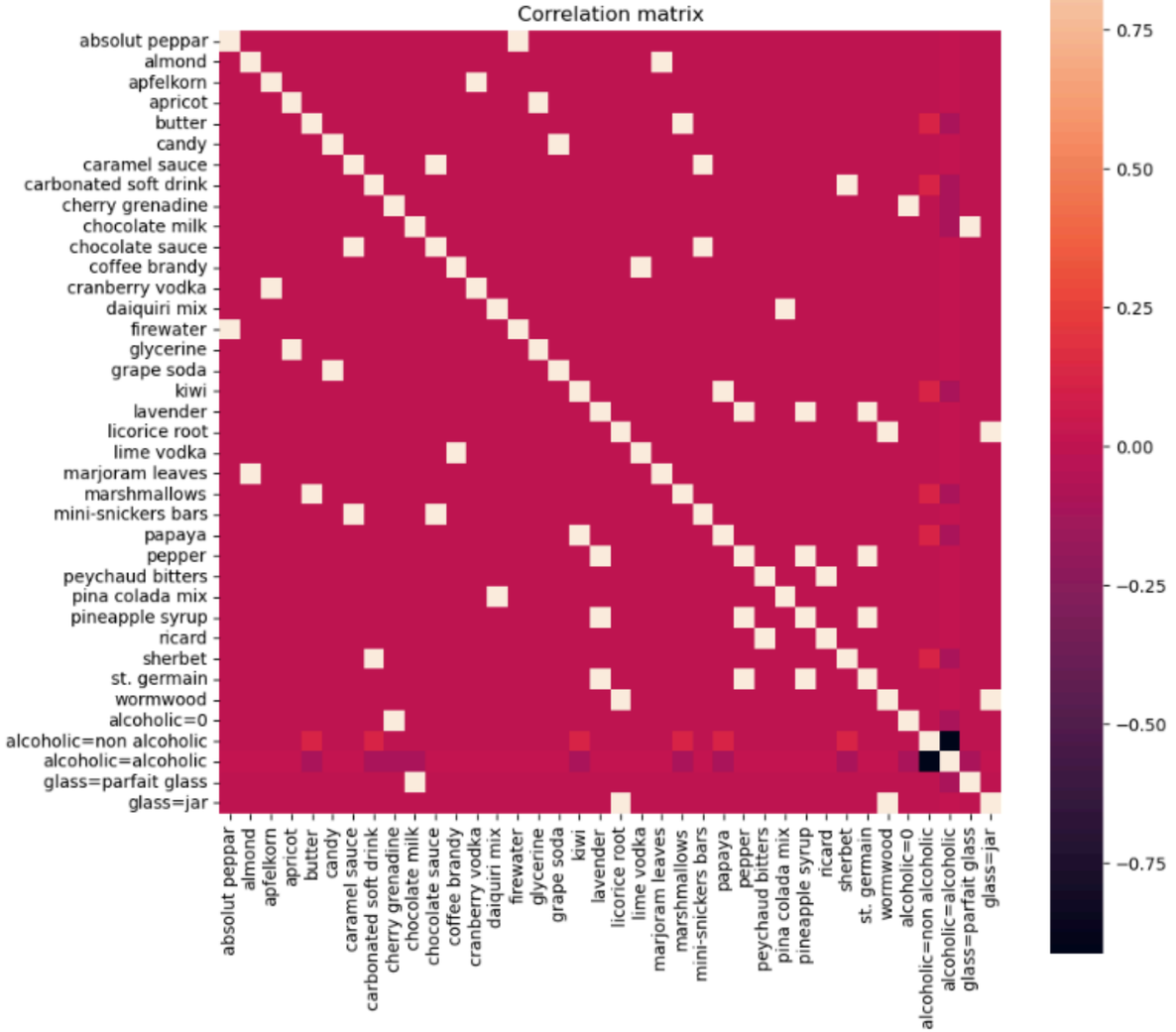
```
gin
0    461
1     85
Name: count, dtype: int64
```

ABBIAMO 85 BEVANDE CON GIN COME BASE

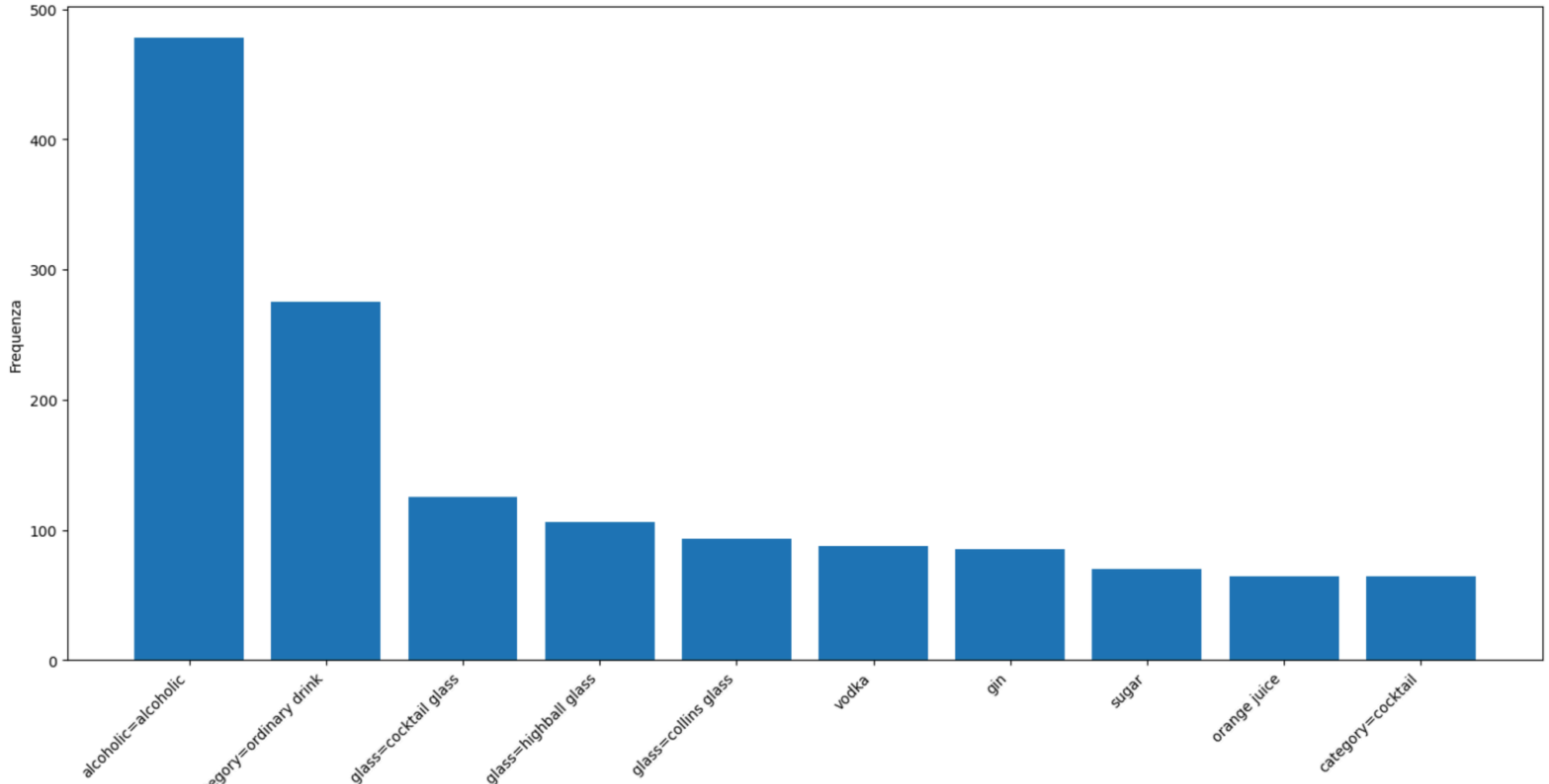
CORRELAZIONE TRA GLI INGREDIENTI

Ho scelto di limitare l'analisi della correlazione per valori di Spearman oltre lo 0.90 in valore assoluto per rendere la heat map più leggibile

Si può notare la correlazione negativa tra alcolici e non alcolici
C'è una correlazione tra il Wormwood (tipologia di assenzio) e le radici di liquirizia



ATTRIBUTI PIÙ FREQUENTI



CLUSTERING: K-MEANS

1

DISTANZA: 1- SMC

Essendo che gli attributi analizzati sono binomiali categorici, ho trovato interessare applicare un algoritmo k-means sostituendo la distanza euclidea con **'1-Simple Matching Coerfficient'**

2

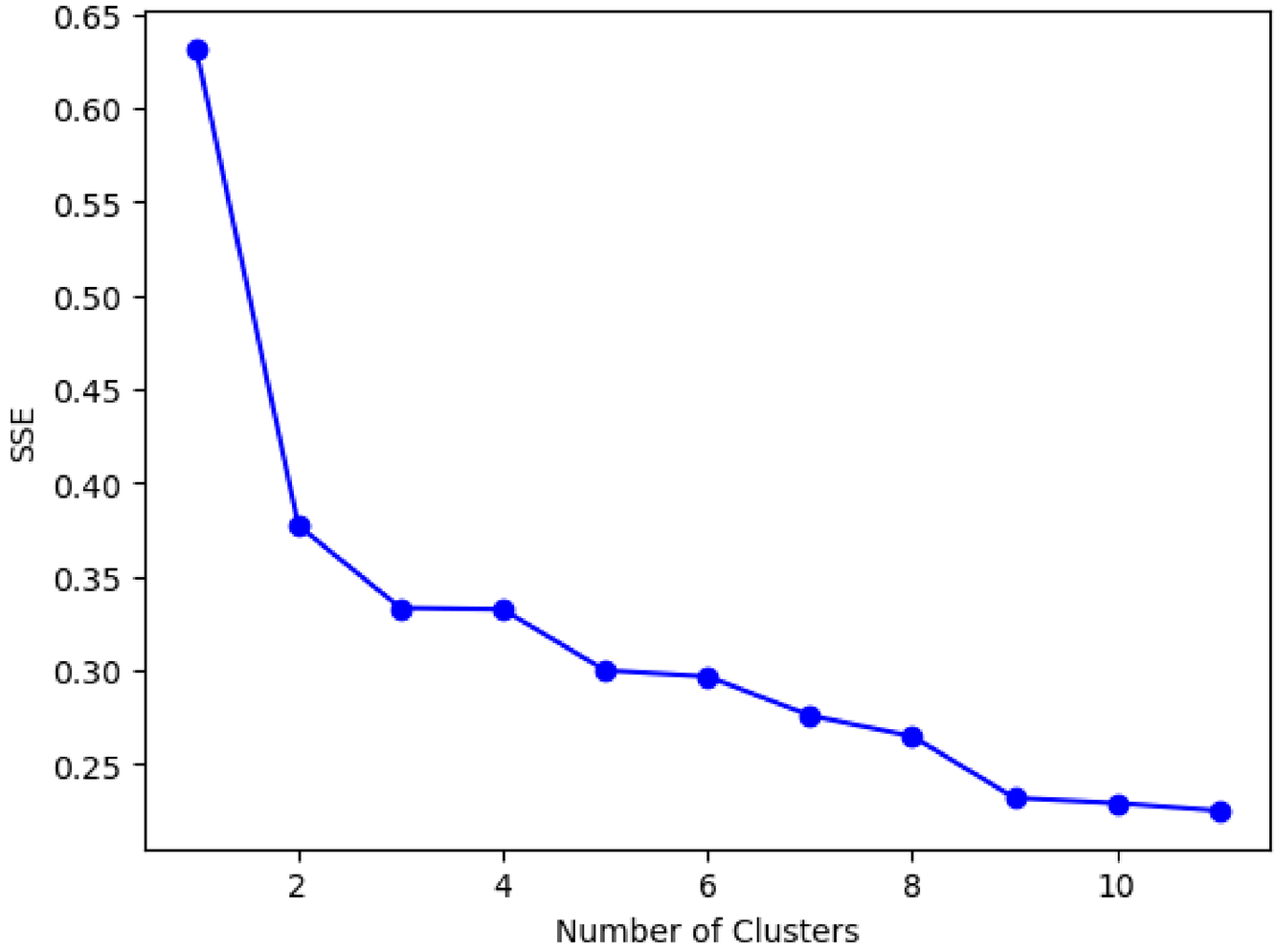
N° CLUSTER=9

Attraverso la analisi dell'SSE all'aumentare del n° dei cluster, sono arrivato alla conclusione di utilizzare **9 come valore di k**:
da quel valore in poi la discesa dell'SSE sembra stabilizzarsi

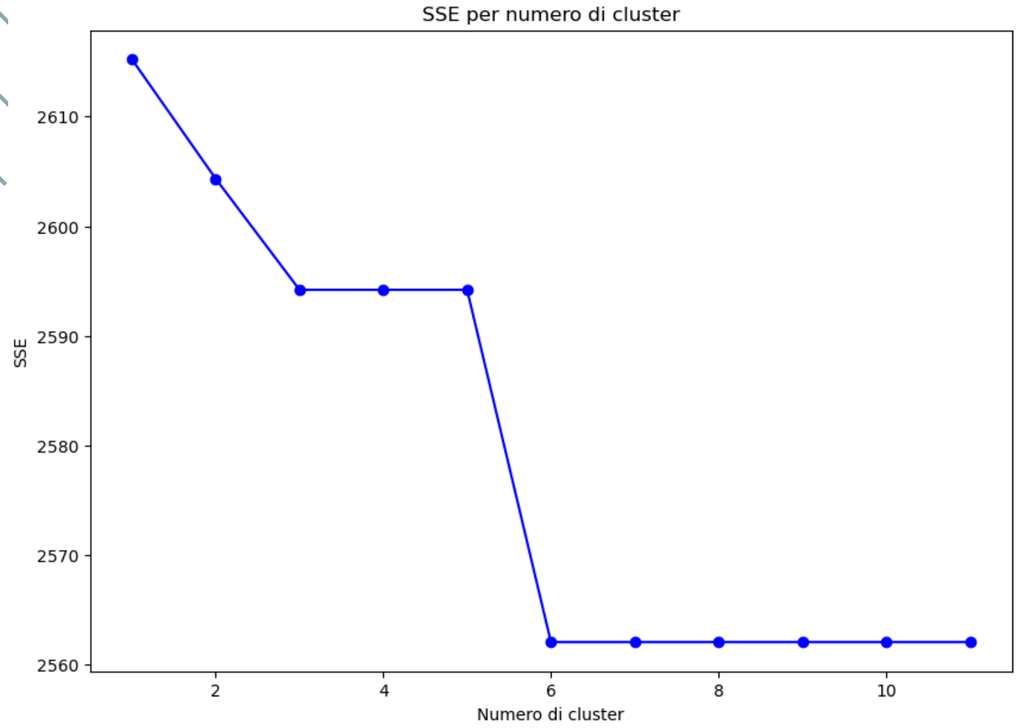
3

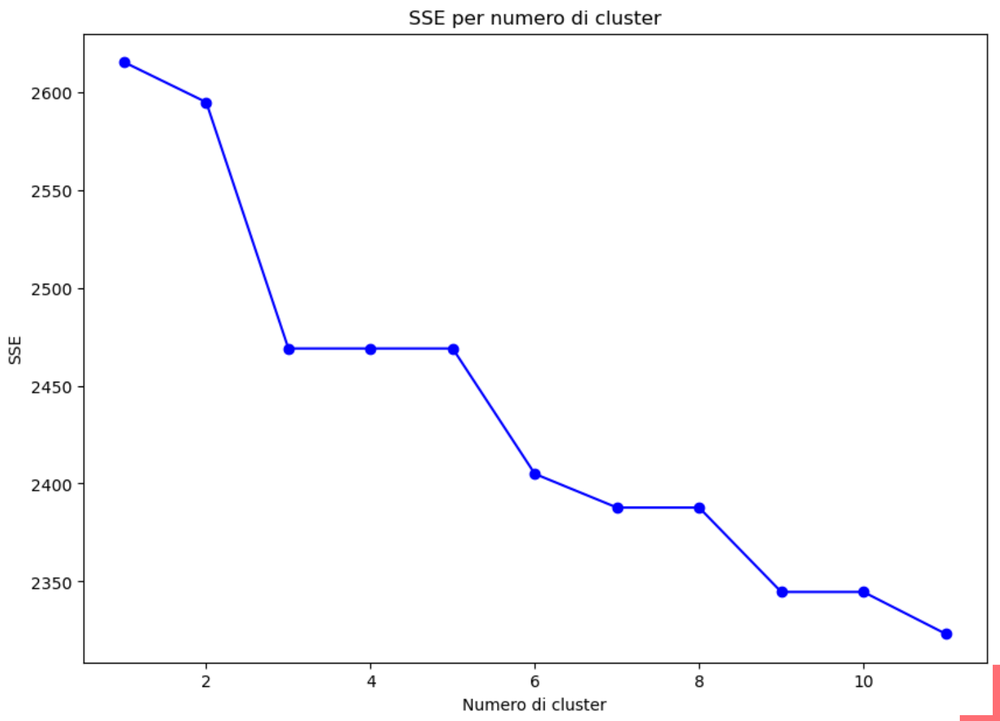
SILHOUETTE COEFFICIENT

Il coefficiente di silhouette per questa clusterizzazione risulta essere circa **0.028**



CLUSTERING GERARCHICO





SINGLE LINK:
N° Cluster: 6
Silhouette Coefficient: 0.15

COMPLETE LINK:
N° Cluster: 9
Silhouette Coefficient: 0.026

ASSOCIATION ANALYSIS

	support	itemsets
0	0.875458	(alcoholic=alcoholic)
1	0.503663	(category=ordinary drink)
2	0.228938	(glass=cocktail glass)
3	0.498168	(category=ordinary drink, alcoholic=alcoholic)
4	0.223443	(glass=cocktail glass, alcoholic=alcoholic)

Algoritmo Apriori

	support	itemsets
0	0.875458	(alcoholic=alcoholic)
1	0.503663	(category=ordinary drink)
2	0.228938	(glass=cocktail glass)
3	0.498168	(category=ordinary drink, alcoholic=alcoholic)
4	0.223443	(glass=cocktail glass, alcoholic=alcoholic)

FPGrowth

I DUE ALGORITMI DANNO GLI STESSI RISULTATI

ASSOCIATION ANALYSIS

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(category=ordinary drink)	(alcoholic=alcoholic)	0.503663	0.875458	0.498168	0.989091	1.129798	0.057233	11.416361
1	(alcoholic=alcoholic)	(category=ordinary drink)	0.875458	0.503663	0.498168	0.569038	1.129798	0.057233	1.151695
2	(glass=cocktail glass)	(alcoholic=alcoholic)	0.228938	0.875458	0.223443	0.976000	1.114845	0.023018	5.189255

Le due regole con più rilevanza nel dataset sono (category=ordinary drink) --> (alcoholic= alcoholic) e (alcoholic= alcoholic)-->(category=ordinary drink).

La prima di queste ha una confidenza molto alta:

La probabilità che una bevanda alcolica sia un ordinary drink è dello 0.98 (98%).

CLASSIFICAZIONE: DISTRIBUZIONE CATEGORIE

alcoholic=alcoholic	glass=cocktail glass	category	beer	cocktail	cocoa	coffee / tea	homemade liqueur	milk / float / shake	ordinary drink	other/unknown	punch / party drink	shot	soft drink / soda
0	0	0	0	6	9	8	0	5	3	22	12	0	0
	1	0	1	0	0	0	1	0	0	0	0	1	
1	0	13	28	0	17	12	11	182	12	23	49	9	
	1	0	29	0	0	0	0	90	0	2	0	1	

Si ha un'alta concentrazione di Ordinary Drink, soprattutto di quelli alcolici non serviti in un Cocktail Glass

In generale, si nota una netta predominanza di alcolici rispetto ad altre tipologie

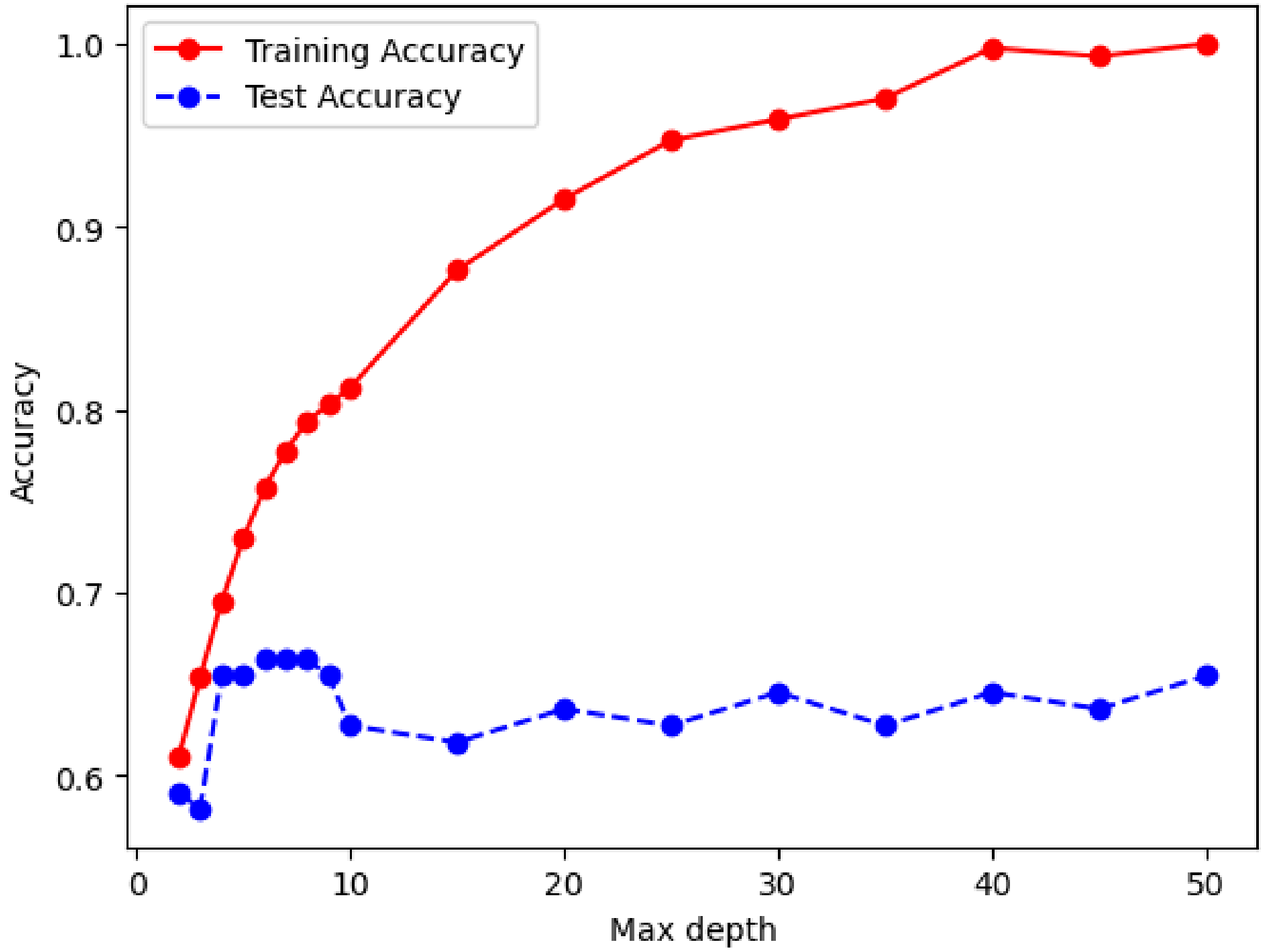
Si denota che in questo dataset non ci sono birre analcoliche: si potrebbe pensare di aggiungerne alcune per un'analisi futura

DECISION TREE

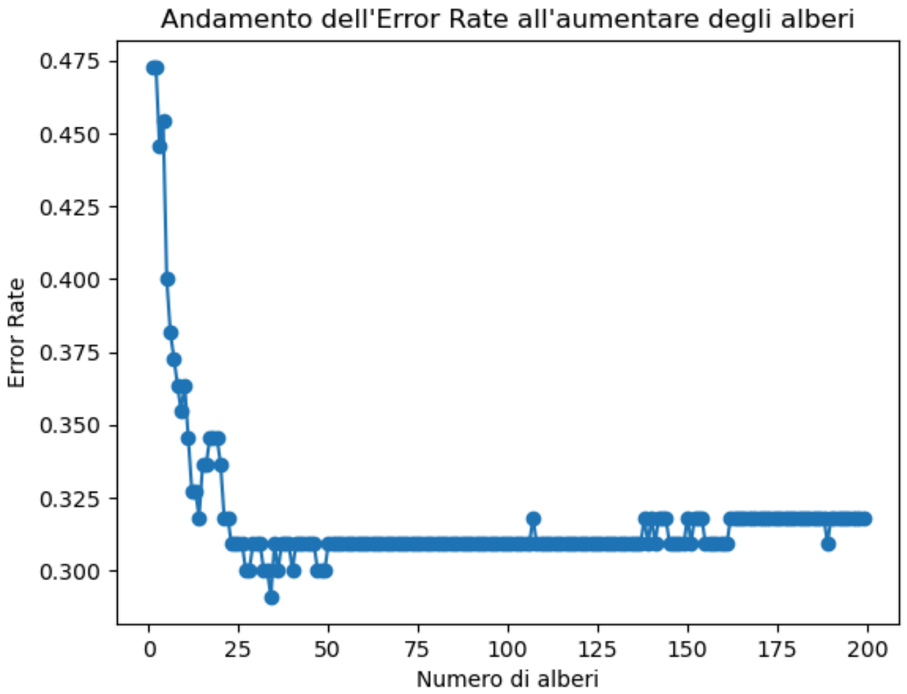
HO DIVISO IL DATASET IN *80% TRAIN E 20% TEST*

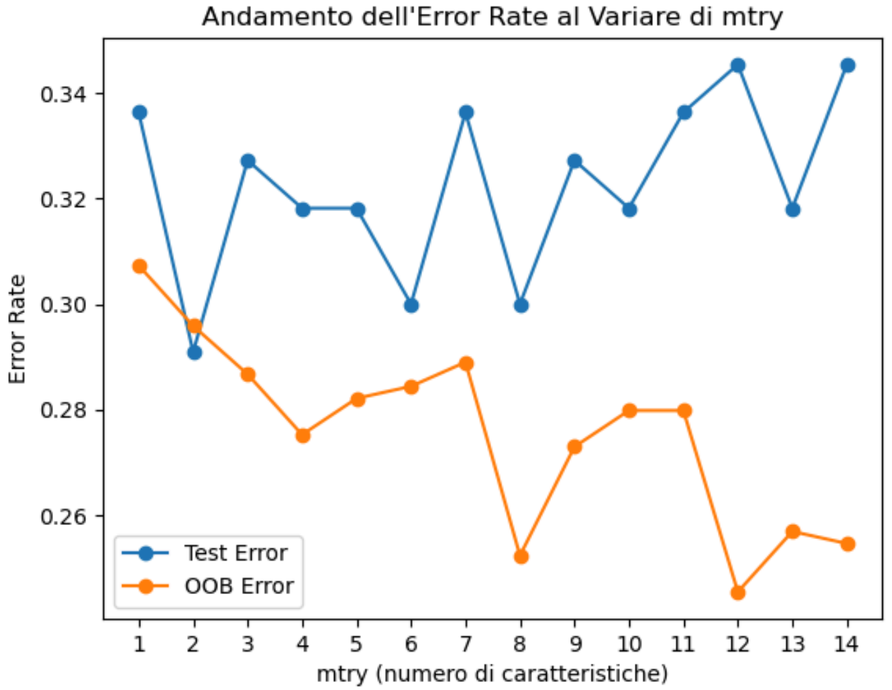
ACCURACY PER IL TEST SET SMETTE DI AUMENTARE CON MAXDEPTH=9

ACCURACY PREDIZIONI: 0,80 PER IL TRAIN E 0,65 PER IL TEST



RANDOM FOREST





ACCURACY MODELLO:
TRAIN: 1.0

TEST: 0.71

L'Error Rate all'aumentare del n° di alberi nella Random Forest si stabilizza a circa 50

Tenendo fisso il numero di albero a 50, trovo sensato prendere 8 come numero di variabili casuali da selezionare come candidati a ogni split in un albero decisionale (mtry).

GRAZIE PER L'ATTENZIONE