



# Flavors

**Federico Zamberlan, Ciro Zar y Lucía Bravo**

**Laboratorio de datos - Cátedra Tagliazucchi - 1er cuatrimestre 2021**

# Esquema de la charla

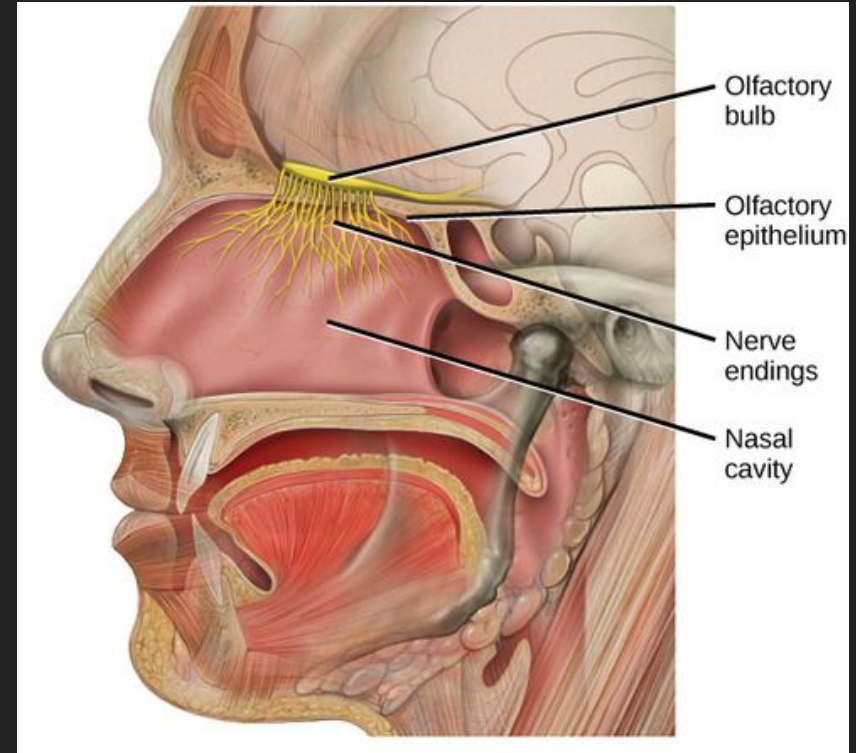
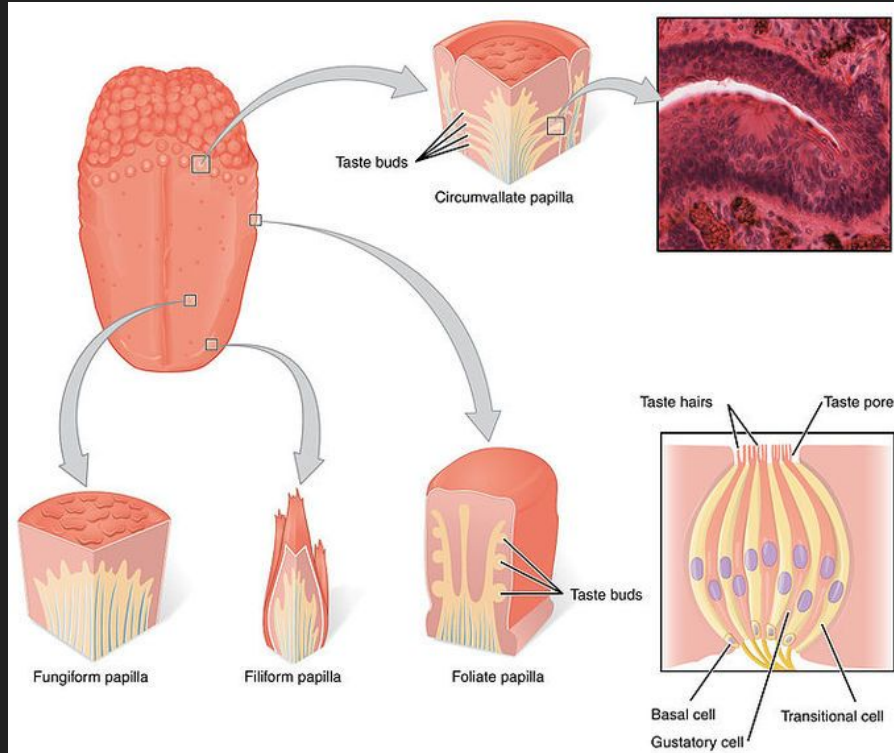
- Motivación
- Visualización y limpieza de datos
- Modelos y resultados
- Conclusiones





# Motivación

¿Qué es un flavor? ¿Qué determina la percepción de un flavor?  
¿Es posible predecir la percepción que se tendrá de una sustancia?



La base de datos empleada fue extraída de [FlavorDB](#)



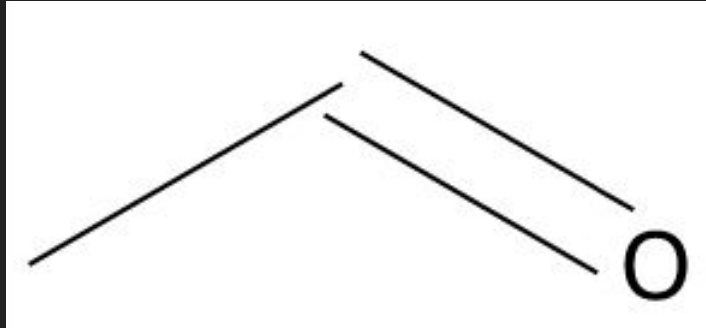
FlavorDB es un repositorio que ofrece información acerca del origen, los flavors, las propiedades fisicoquímicas, entre otras cosas, de moléculas.

Es una herramienta útil para la industria alimenticia.

Tiene 25595 moléculas de flavor, de las cuales 2254 están asociadas a 936 ingredientes naturales.

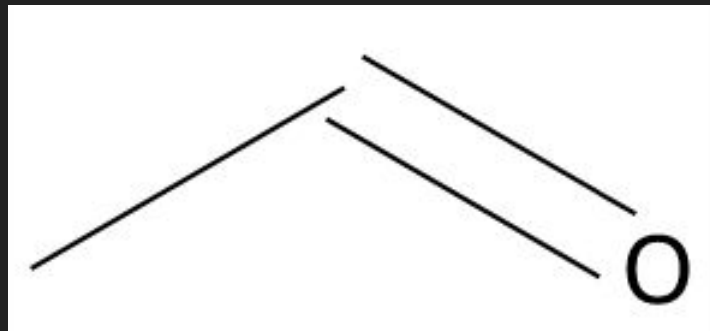


## Acetaldehído



Tags	Ethereal	Pungent
	Ether	Fruity
	Whiskey	Aldehydic

Acetaldehído



Tags

Ethereal

Ether

Whiskey

Pungent

Fruity

Aldehydic



# Visualización y limpieza de datos

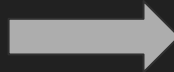




Citric {  
Citric  
Citric peel  
Citrus  
Citrus peel

Milk {  
Milk  
Milky  
Dairy  
Hot milk

25595 moléculas  
700 flavors

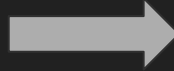


25106 moléculas  
523 flavors

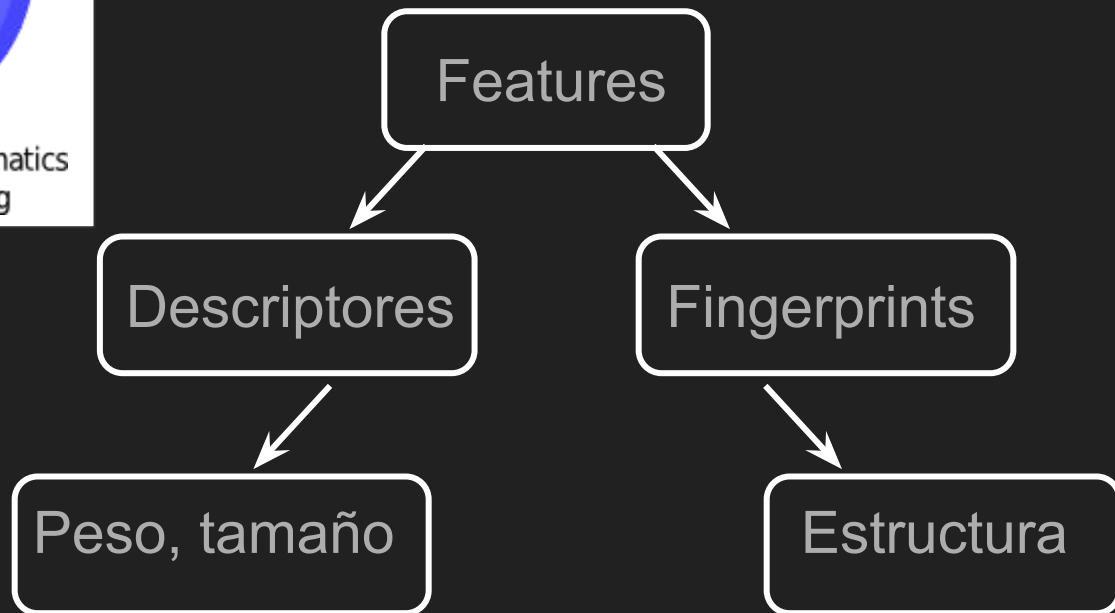
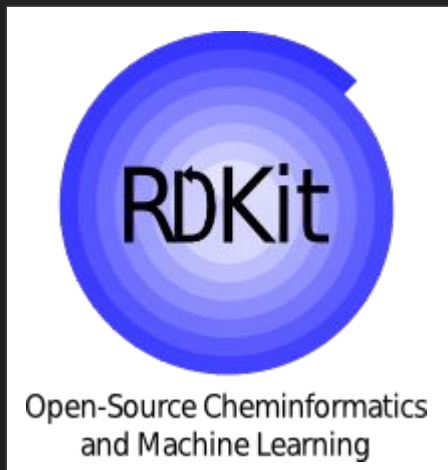
Se eliminaron flavors menos frecuentes:

163 flavors aparecían una sola vez, 70 flavors dos veces, etc

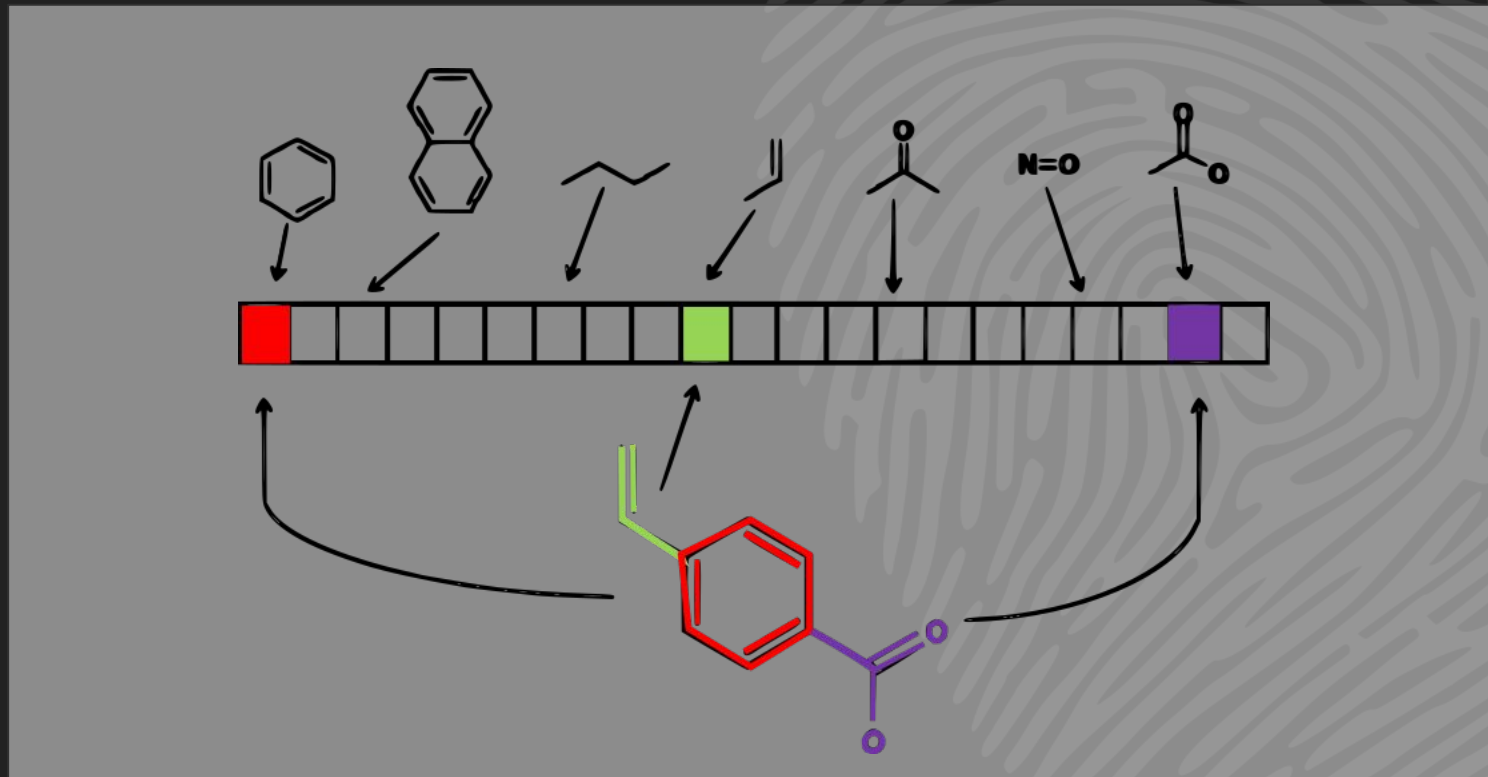
25106 moléculas  
523 flavors



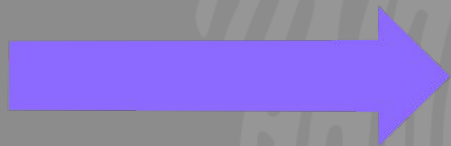
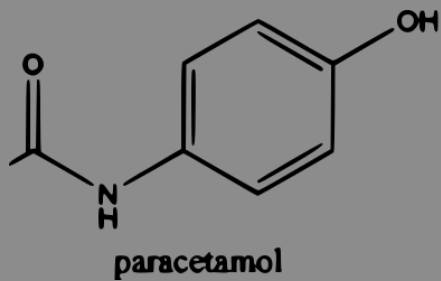
25106 moléculas  
135 flavors



# Molecular Fingerprint



# Molecular Fingerprint



(0,0,1,0,1,0,0,...,0,1,0,0)



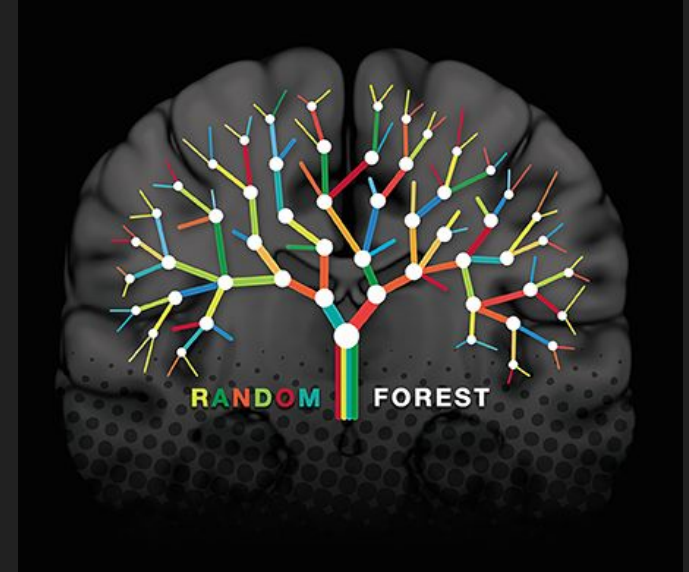
# Modelo I - Random Forest Classifier

Tag: Bitter

Folds: 10 (Stratified)

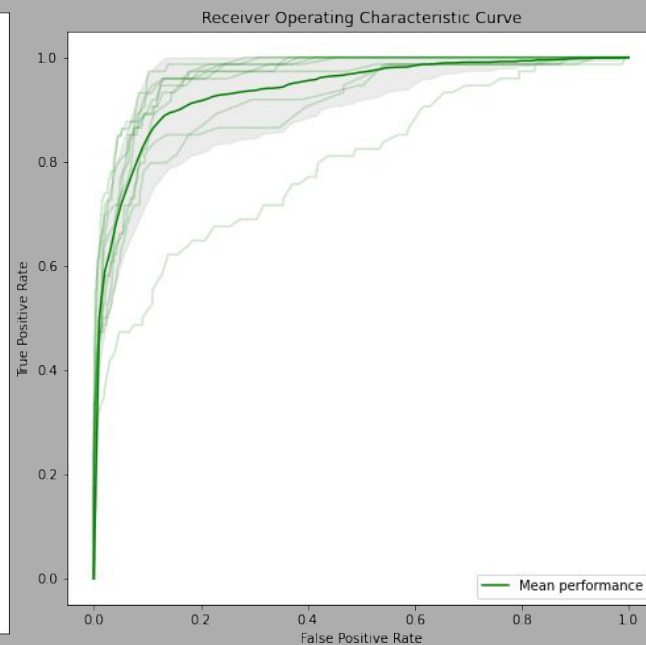
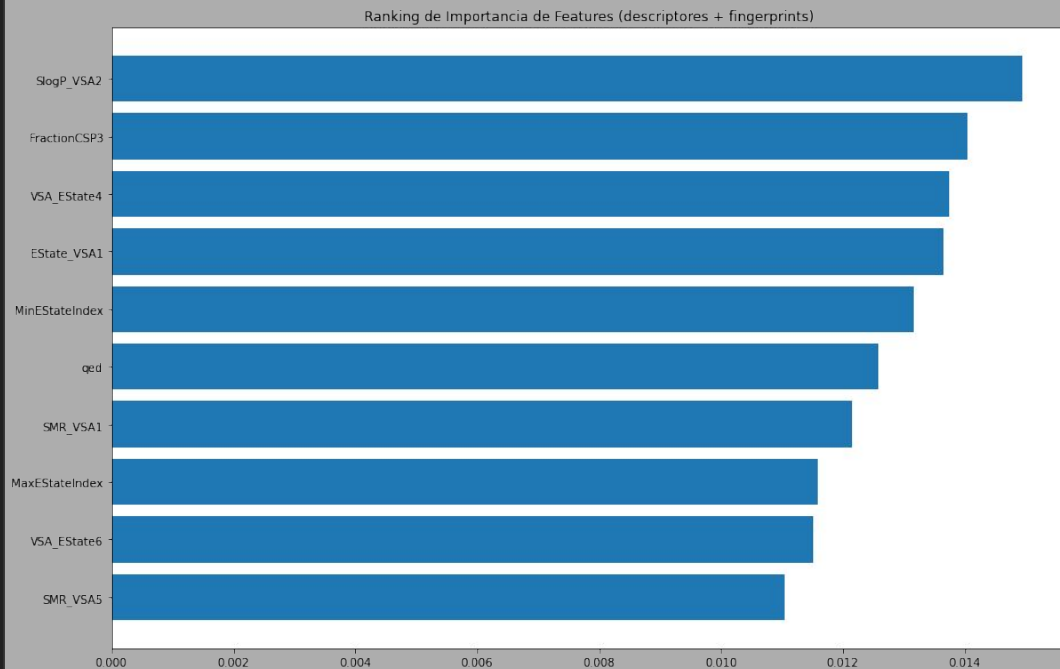
Features:

- Fingerprints & Descriptores combinados
- Solo Descriptores
- Solo Fingerprints



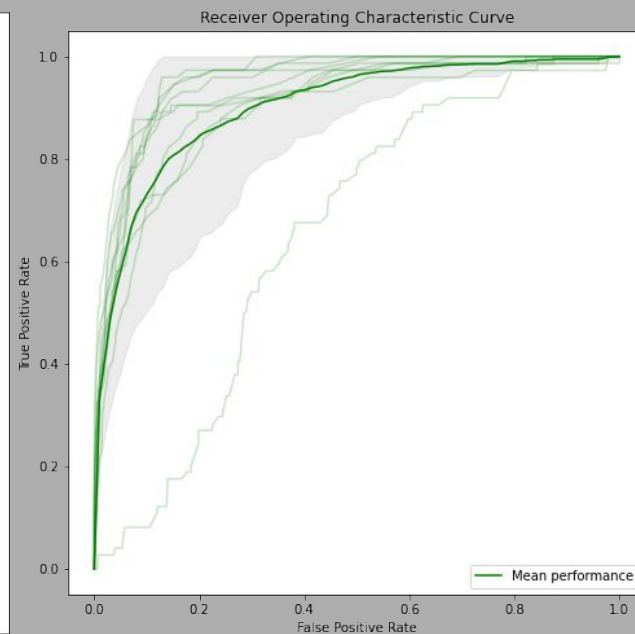
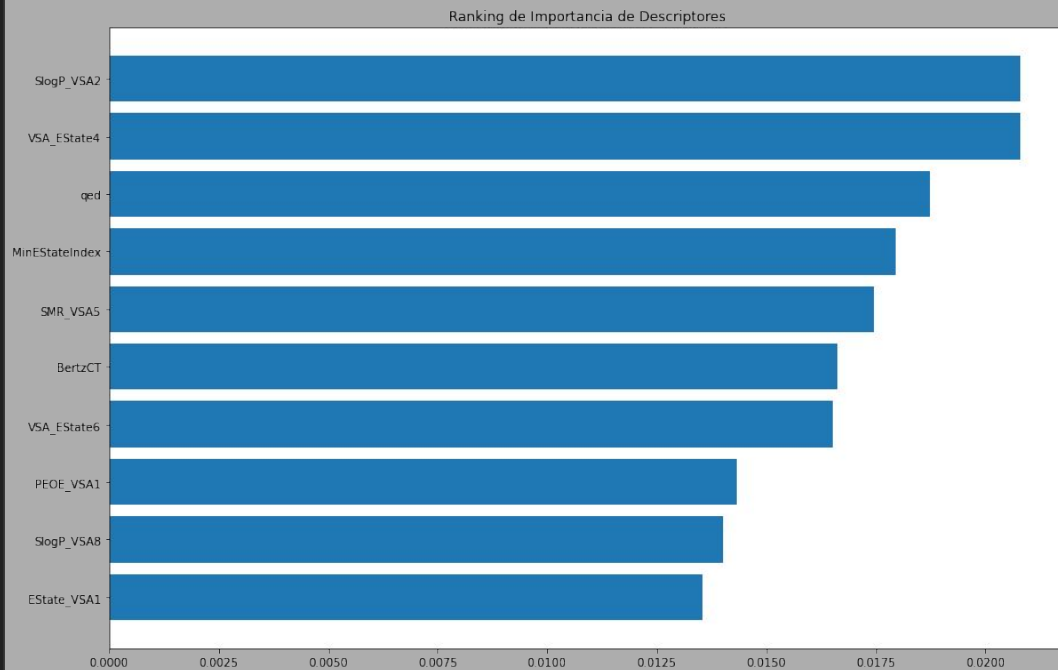
# Fingerprints + Descriptores

AUC promedio= 0.936



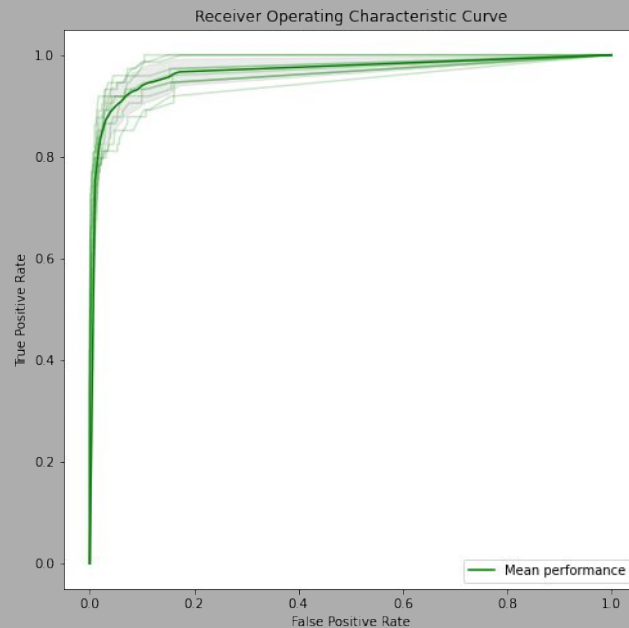
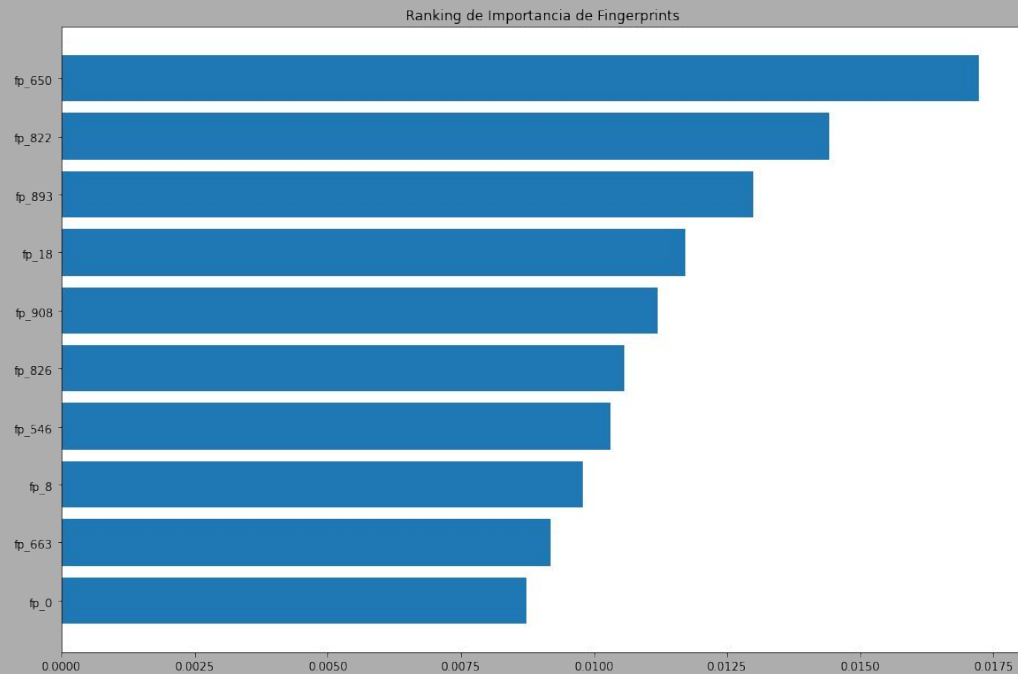
# Descriptores

AUC promedio= 0.901

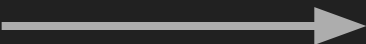


# Fingerprints

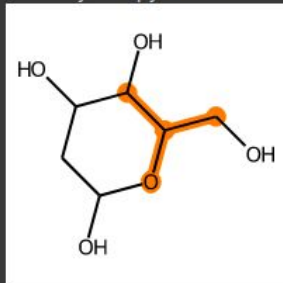
AUC promedio= 0.96955



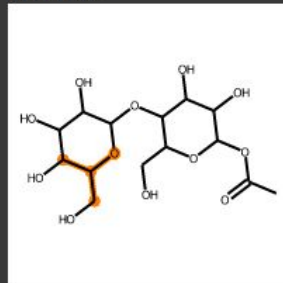
# Visualización - Fingerprints

Para bitter  fingerprint 650

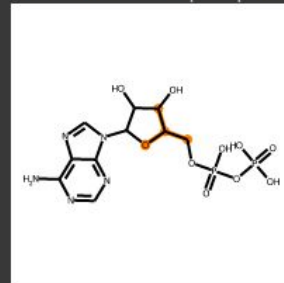
2-Deoxyhexopyranose



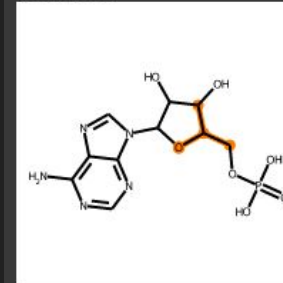
AC1L180C



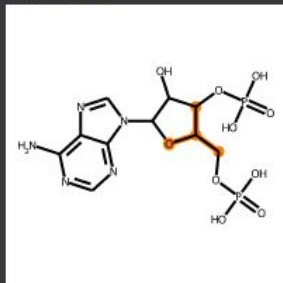
Adenosine-5'-Diphosphate,



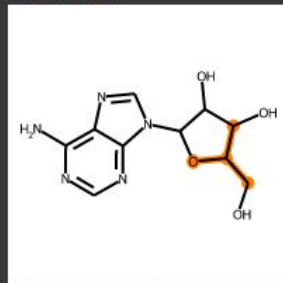
NSC20264



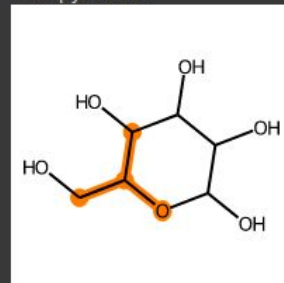
AC1L18F4



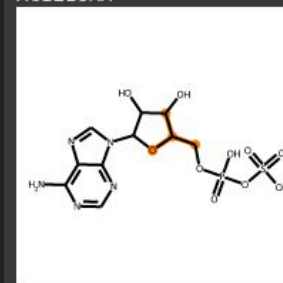
3228-71-5



Hexopyranose



AC1L18RM





# Modelo II - Clustering

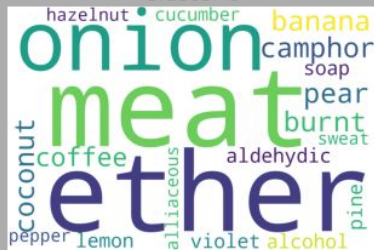
Clusters: 4

Método: PCA (100 componentes) + KMeans

Features: Fingerprints & Descriptores combinados

# Modelo II - Clustering

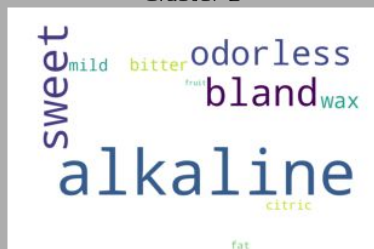
Cluster 0



Cluster 1



Cluster 2

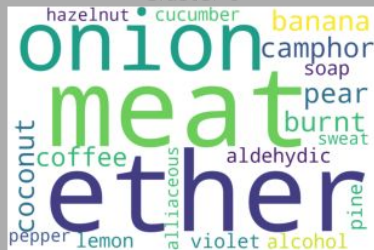


Cluster 3



# Modelo II - Clustering

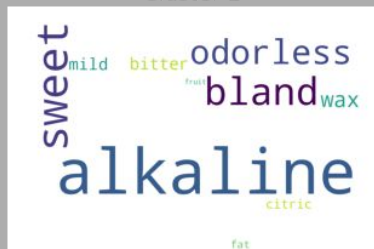
Cluster 0



Cluster 1



Cluster 2

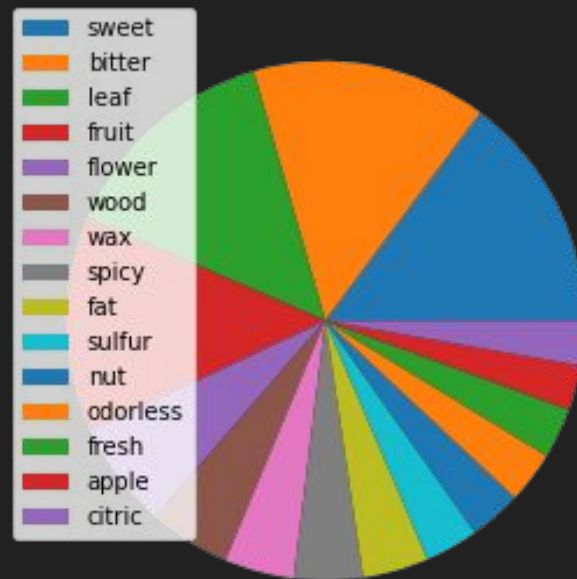
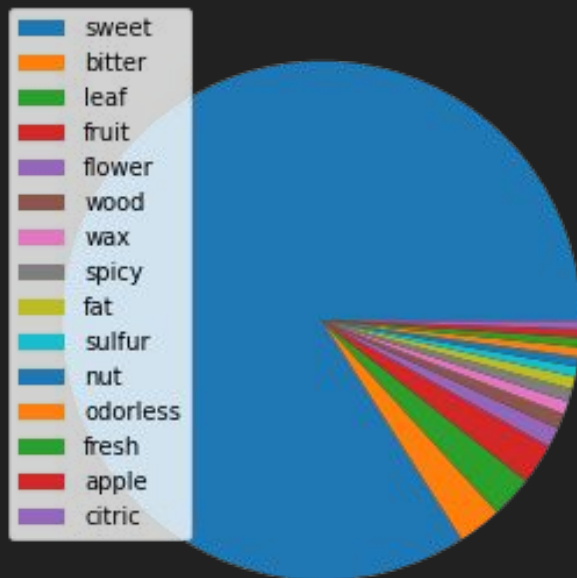


Cluster 3



# Modelo II - Clustering

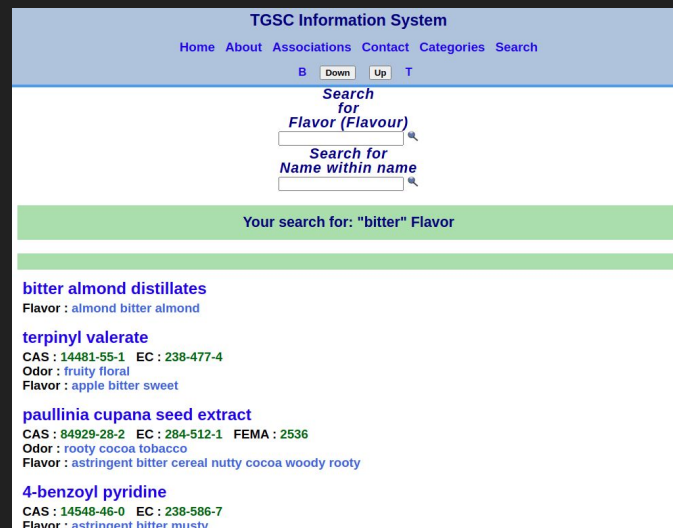
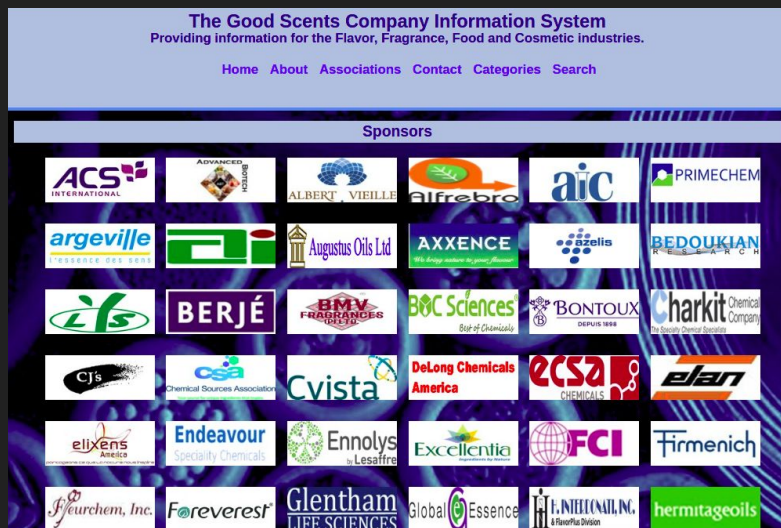
Se removieron moléculas cuyo único tag era sweet



# Word2Vec

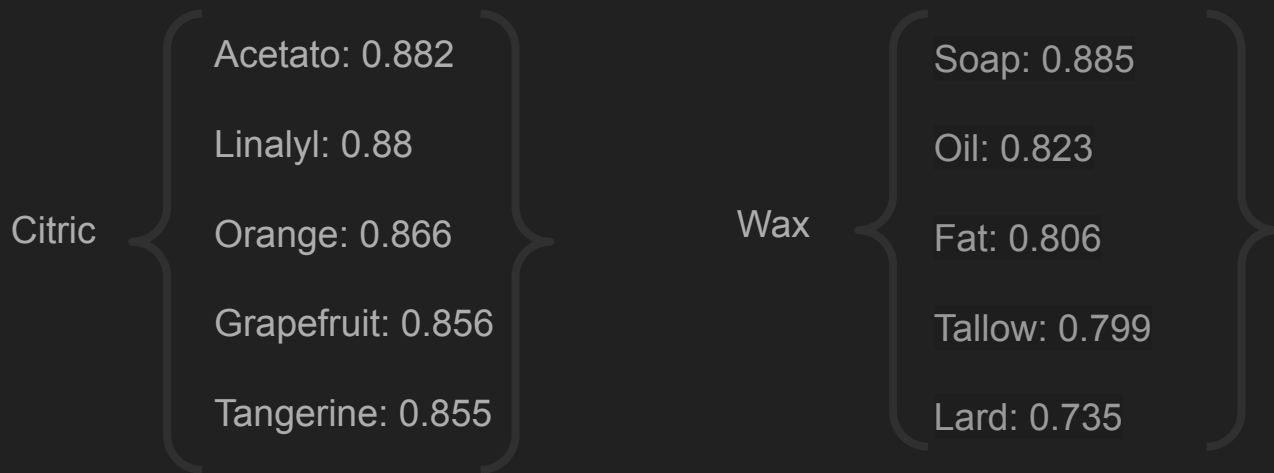
Criterio para analizar la cercanía entre elementos de un clusters y entre clusters:

Se empleó la base de datos [The Good Scent Company](#) creada en 1994.





Se entrenó un modelo Word2Vec para calcular la cercanía entre cada uno de esos tags, tomando sólo los que nos interesan.



# Modelo II - Clustering

$C = \text{Similaridad en el cluster} / \text{Similaridad entre clusters}.$

Si  $C < 1$  los clusters son muy similares entre sí. (0.54 en primera aproximación)

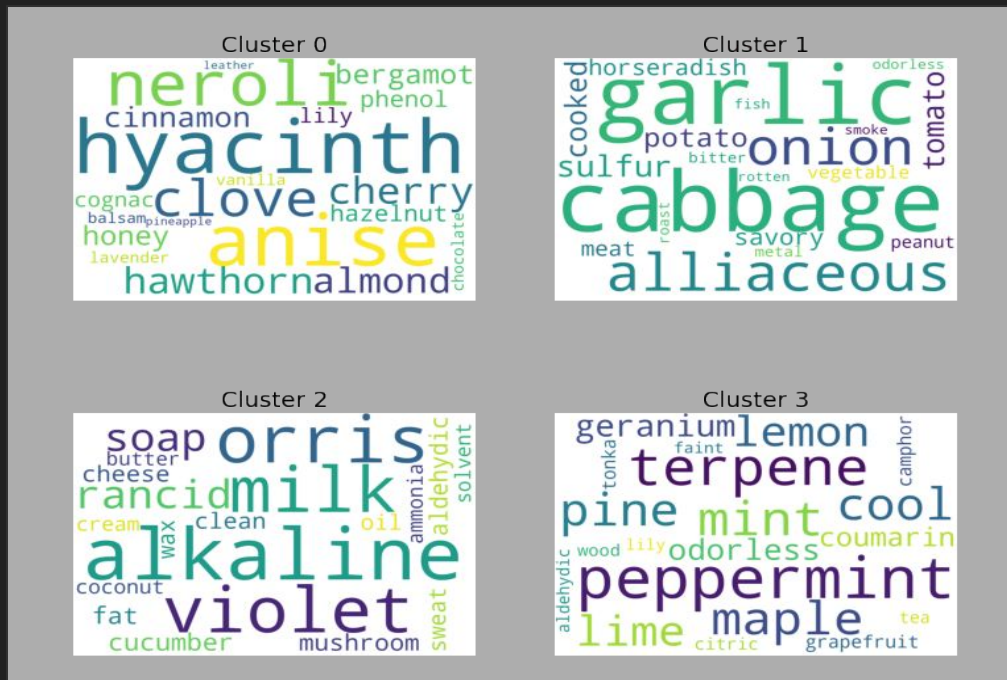
Es importante obtener  $C > 1$ .

# Clusters - Dataframe Balanceado

Cluster 1: 0.768

0 vs 2: 0.034

C = 3.69



# Conclusiones

En cuanto a lo realizado:

- Es posible identificar fragmentos de moléculas importantes para clasificar.
- Se mejoró la clusterización de los tags.
- Se pudo construir un modelo predictivo que lograra entender la experiencia subjetiva basado en características de las moléculas.

# Conclusiones

Algunas ideas de lo que es posible hacer

- Ver si para tags como “aldehídico” el feature importance muestra en la estructura el grupo funcional aldehído.
- Ampliar la base de datos y analizar además de flavors, toxicidad de las moléculas.
- Intentar predecir el gusto final de una comida a partir de los ingredientes individuales de la receta



¡Gracias!