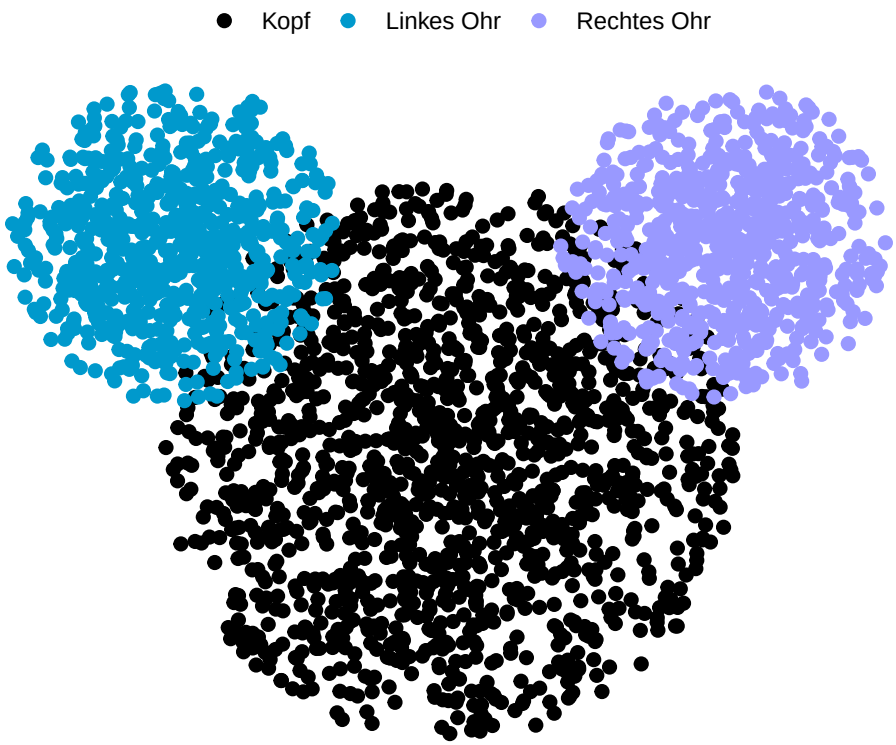
Praktikum 2: EM-Algorithmus

Zuerst schauen wir uns die Daten an. Durch einen einfachen Scatterplot wird, warum der Datensatz den Namen “maus” trägt.

Im nächsten Schritt schauen wir und auch die Klassenverteilung an.



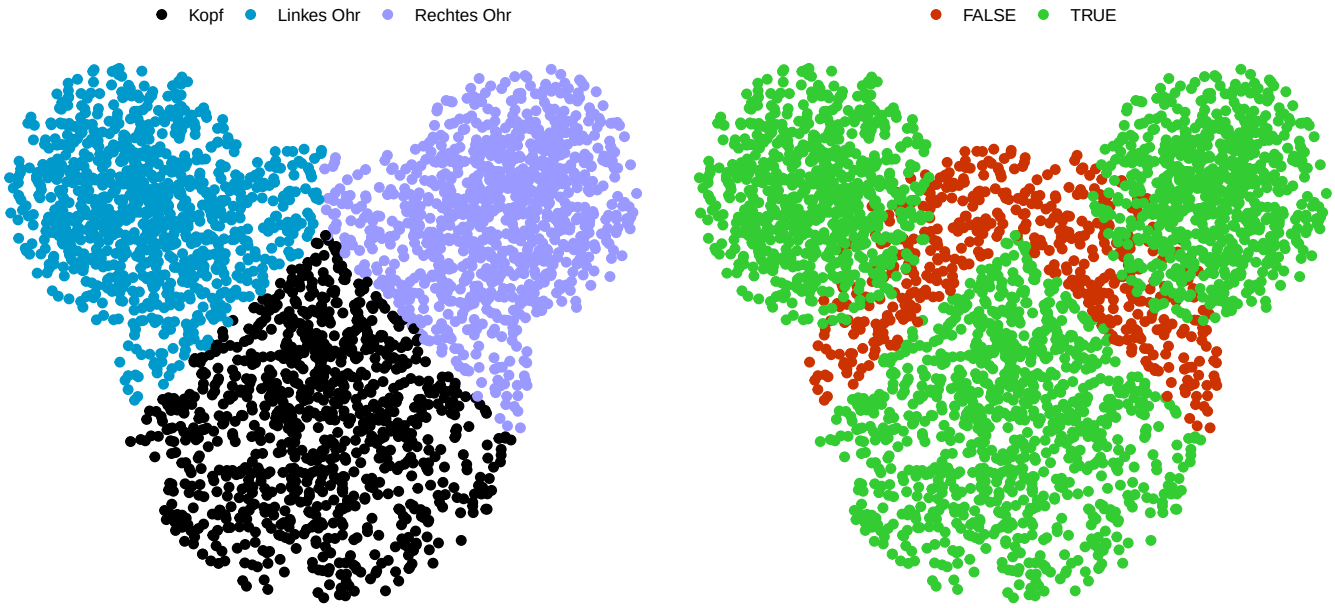
Es gibt 3 Klasse:

Tabelle 1: Datenverteilung

Klasse	Summe
Kopf	1482
Linkes Ohr	739
Rechtes Ohr	731

Wir stellen fest, dass der Schnitt der knovexen Hüllen der Ohren mit dem Kopf nicht leer ist. Da sowohl kmeans als auch EM jedem Punkt nur eine Klasse zuweisen und Überschneidungen der konvexen Hüllen vermeiden, sorgt dies direkt für einen gewissen unvermeidbaren Fehler. Auch nach Augenmaß und der Tabelle 1 ist festzustellen, dass die Dichte der Datenpunkte in den Ohren jeweis höher als im Kopf ist.

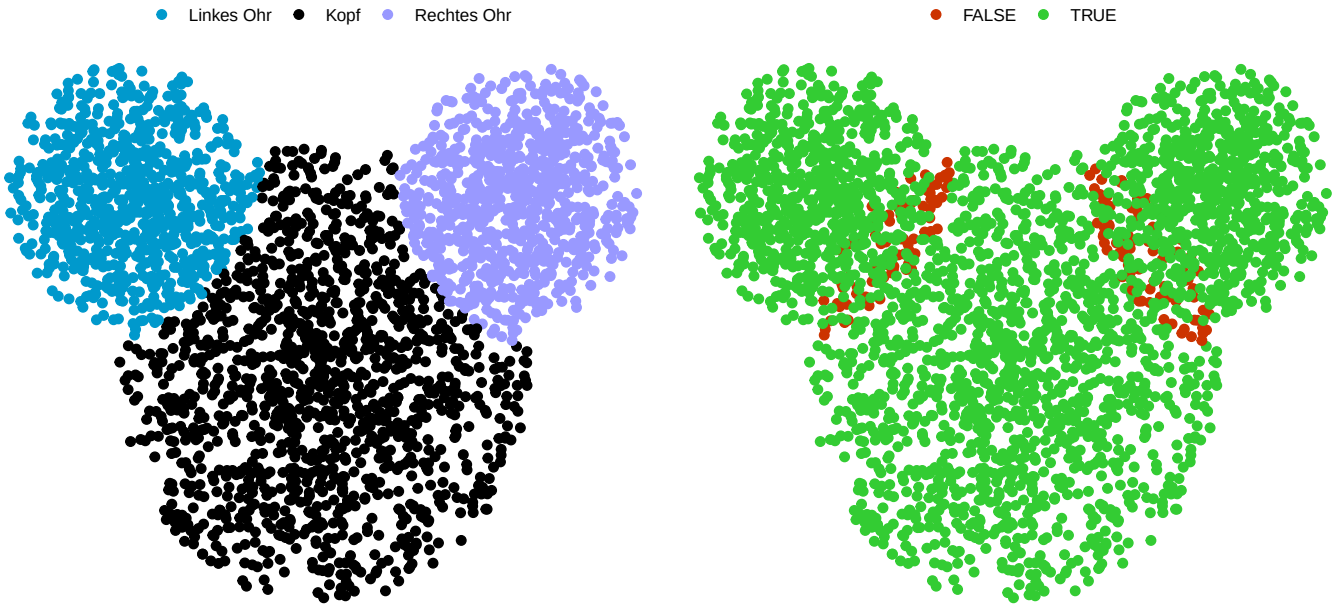
Nun berechnen wir ein Clustering mittels kmeans-Algorithmus. In der linken Grafik sieht man einen Scatterplot des Clusterings und in der rechten Grafik einen Scatterplott mit Klassifizierung nach “TRUE” und “FALSE” je nachdem, ob der Algorithmus richtig klassifiziert hat.



Accuracy= 0.824
Silhouette−Coefficient= 0.504

Wir erhalten eine Genauigkeit von 0.824. Was man direkt sehen kann ist, dass kmeans Punkte gleichmäßig Centroiden nach der euklidischen Distanz zuweist. Dadurch werden einige Punkte des Kopfes falsch zugewiesen, dafür sämtliche Punkte der Ohren korrekt klassifiziert. Das Ergebnis ist in Ordnung, aber nicht besonders gut.

Nun berechnen wir ein Clustering mittels EM-Algorithmus. In der linken Grafik sieht man einen Scatterplot des Clusterings und in der rechten Grafik einen Scatterplott mit Klassifizierung nach “TRUE” und “FALSE” je nachdem, ob der Algorithmus richtig klassifiziert hat.



Accuracy= 0.942
Silhouette−Coefficient= 0.496

Wir erhalten eine Genauigkeit von 0.942. Was man direkt sehen kann ist, dass die Zuweisung der Klassen visuell deutlich besser als bei kmeans ist. In Zahlen erkennt man das auch daran, dass die Genuigkeit höher ist. Da der EM-Algorithmus anhand der bedingten Wahrscheinlichkeit klassifiziert wird nochmals klar, dass die Dichte der Datenpunkten in der Ohren jeweils höher als beim Kopf ist und der EM-Algorithmus deswegen ein besseres Ergebnis liefert.

Als Ergebnis erhalten wir, dass der EM-Algorithmus ein besseres Ergebnis liefert.

Tabelle 2: Ergebnisse

Algorithmus	Genauigkeit	Silhoutte.Koeffizient
kmean	0.824	0.504
em	0.942	0.496

Der Silhoutte-Koeffizient ist nach seiner Definition von der Clusteranzahl unabhängig, aber abhängig davon wie viele Punkte in den einzelnen Clustern liegen und wie scharf diese voneinander getrennt sind. Der Koeffizient wird genutzt, um bei Algorithmen, die mit unterschiedlichen Startpunkten zu unterschiedlichen lokalen Maxima laufen können, das Clustering zu bewerten. Sprich umso höher der Silhouetten-Koeffizient, umso stärker sind die Cluster strukturiert und umso schärfer sind diese getrennt. Dies ist bei beiden Algorithmen der Fall und man könnte den Koeffizienten damit nutzen die Parameter für das jeweilige Clustering zu optimieren. Aber der Silhoutten-Koeffizient basiert auch auf die euklidische Distanz, wodurch die Anwendung auf den EM-Algorithmus damit nicht so viel Sinn ergibt, da für die Maßzahlen jeweils unterschiedliche Maße vorliegen. Deswegen bietet sich zum Vergleich zwischen den Algorithmen eher die Genauigkeit an.

Zusammenfassung der Plots als Gesamtergebnis

Die Plots auf der linken Seite gehören zum kmeans-Algorithmus und die auf der rechten Seite vom EM-Algorithmus.