הריצו מודל K-Means )על סט האימון( עם ערכי ברירת המחדל )"המאשכל הבסיסי"(,

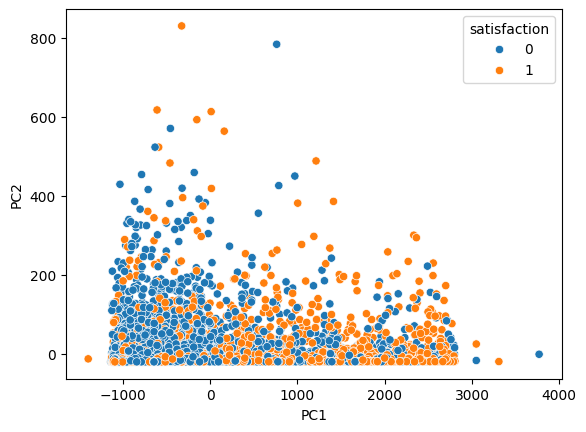
פרט להגדרת מספר האשכולות, אותו תגדירו כך שיתאים לבעיה עמה אתם מתמודדים

ועבורה ידוע לכם מספר המחלקות.

נבחר מספר אשכולות k = 2, מכיוון שאנו יודעים שיש לנו שני מחלקות של שביעות רצון (שביעות רצון, חוסר שביעות רצון).

תמונה שמכילה טקסט, תרשים, צילום מסך, מקביל

התיאור נוצר באופן אוטומטי



. דונו בתוצאות . מה טיב ההתאמה בין האשכולות למחלקות? כיצד שויכ ו תצפיות

לאשכולות? לוו את ההסברים בגרפים מתאימים שיציגו את הנתונים )או חלק מהם(

במרחב. ניתן להשתמש בשיטות להורדת ממד כגון PCA

1. המדד Adjusted Rand Index (ARI) שהתקבל הוא 0.096. מדד ה-ARI נע בין -1 ל-1, כאשר ערך של 1 מצביע על התאמה מושלמת בין שתי החלוקות, וערך של 0 מצביע על כך שהחלוקה היא אקראית לגמרי. ערך שלילי מצביע על כך שהחלוקה גרועה מאקראית.

ערך ה-ARI של 0.096 הוא ערך נמוך מאוד, דבר המצביע על כך שההתאמה בין האשכולות שנוצרו על ידי מודל ה-K-Means לבין המחלקות האמיתיות של שביעות הרצון היא נמוכה.

תמונה שמכילה צילום מסך, טקסט

התיאור נוצר באופן אוטומטי

**לא חובה:** במנותק ממשתנה המטרה של סיפור המקרה, בו יש מספר ידוע של קבוצות,

אמנו כעת מחדש שמונה מודלים של אשכול עם ערכי K משתנים )כלומר, עם מספר

שונה של אשכולות בכל פעם(. מהו ה- K בו תבחרו ? הסבירו את אופן בחירתכם .

האם הערך מתקשר לסיפור המקרה? אם לא, מה יכולה להיות הסיבה לכך?

נשתמש במדדים שונים של הומוגניות והפרדה (silhouette , davies bouldin) . כאשר :

אינרציה היא מדד להומוגניות - סכום המרחקים בריבוע של הדגימות למרכז הצביר הקרוב ביותר שלהן.

ציון דייויס-באולדן ומדד סילואט מערבבים הומוגניות והפרדה.

איכות קיבוץ טובה יותר מיוצגת על ידי אינרציה נמוכה יותר וציון דייויס-באולדן נמוך יותר עם ציון סילואט גבוה יותר.לאחר שיכלול של שלושת המדדים נבחר את מספר האשכולות כך שהוא ייתן תוצאה סבירה שכול המדדים, כלומר אף מדד לא אהיה גרוע ולכן נבחר במספר האשכולות -5

תמונה שמכילה קו, תרשים, עלילה, טקסט

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, קו, תרשים, עלילה

התיאור נוצר באופן אוטומטי

תמונה שמכילה קו, טקסט, תרשים, עלילה

התיאור נוצר באופן אוטומטי

תמונה שמכילה צילום מסך, טקסט

התיאור נוצר באופן אוטומטי