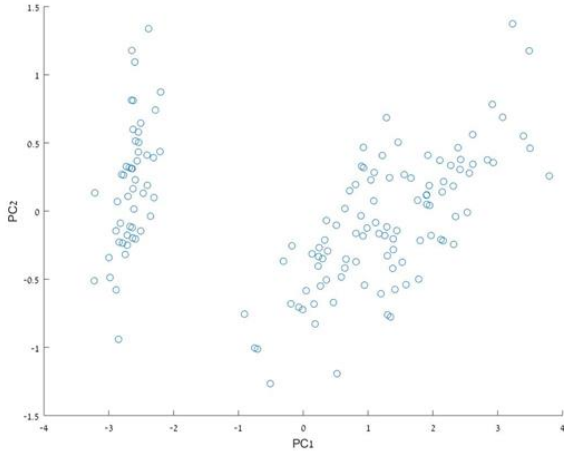


שאלות פתוחות

שאלה פתוחה 3

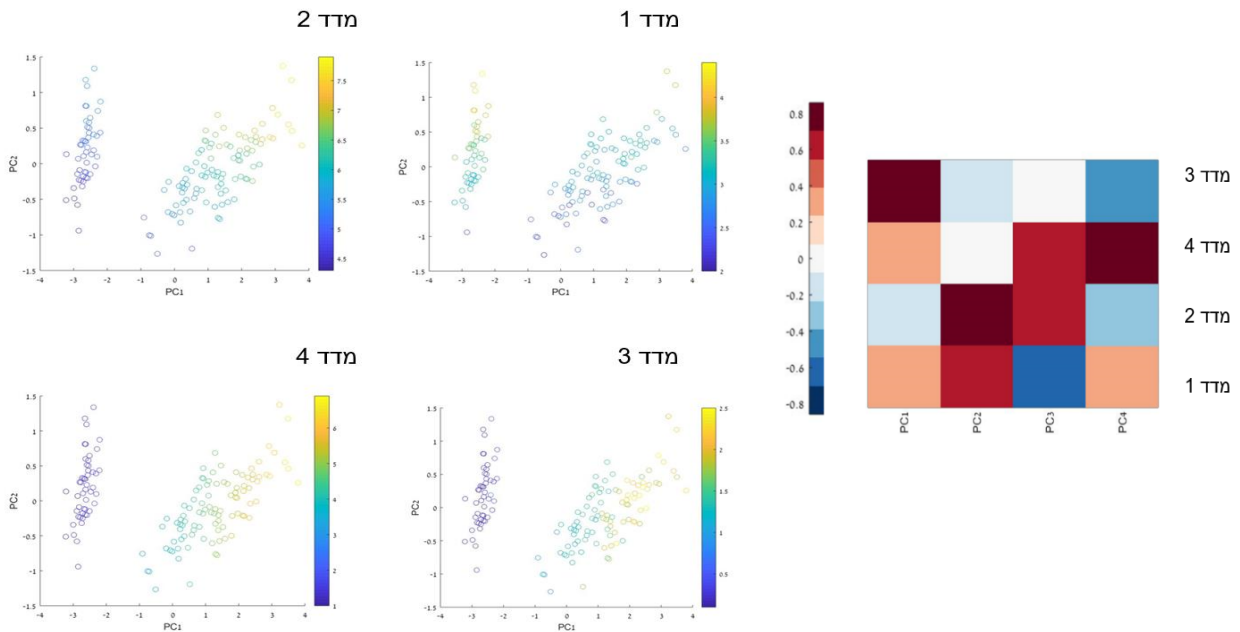
במחקר אודות חולים נמדדו ארבעה מדדים לכל אדם בקבוצה של אנשים חולים וקבוצה של אנשים בריאים (כביקורת). מטרת המחקר היתה למצוא קשר בין אחד המדדים להתקדמות המחלה. בתור התחלה שורטט PCA של כל הנתונים, תוך החלטה שלא לנרמל את המדדים לאותה הסקאלה.



א. לכמה קבוצות מחולקים האנשים? באיזה PC ניתן לראות את ההפרדה?

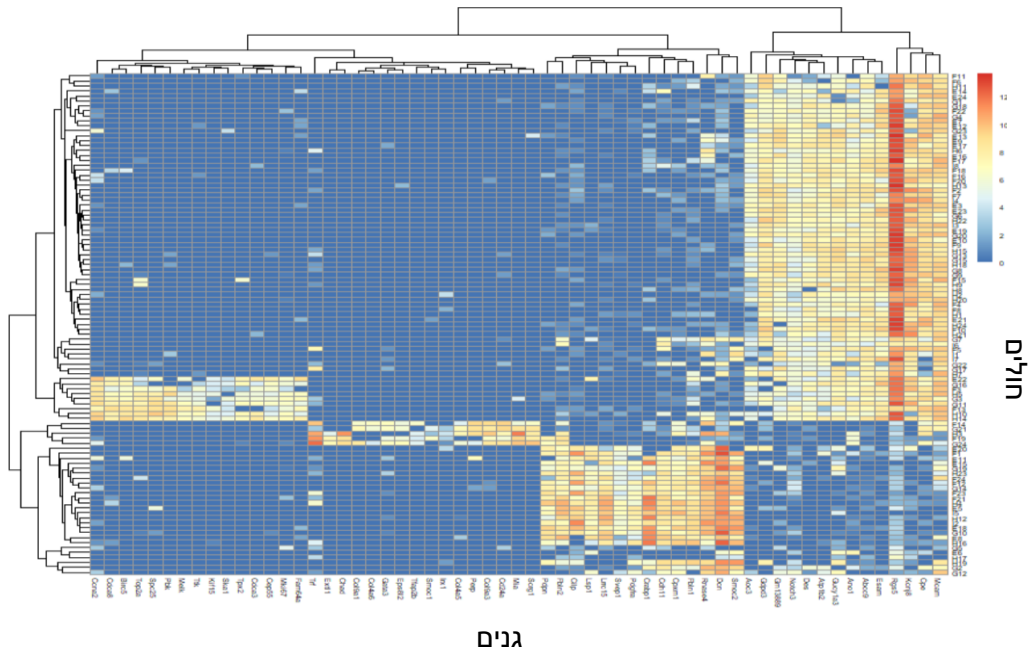
2 קבוצות. PC1.

ב. בשלב הבא שרטטו (1) heatmap של ה-loadings של כל מדד בכל PC, ו בנוסף, על גבי שרטוט ה-scatter של ה-pca צבעו כל דוגמא לפי ערך המדד (שים לב שטווח הערכים לכל מדד שונה). תוך התייחסות גם ל-HEATMAPS וגם ל-SCATTER, בחר פרמטר אחד שהוא הכי משמעותי להפרדה לקבוצות. הסבר.



מדד 3. מדד 3 ו-4 קורלטיבים להפרדה בין הקבוצות כפי שניתן לראות ב-PCA, ומדד 3 הוא בעל הפקטור הגדול ביותר ב-PC1, שם נראית ההפרדה בין הקבוצות, לעומת מדד 4 שהוא משמעותי יותר ב-PC3 וב-PC4 אותם אנחנו לא רואים.

בשלב הבא, החליטו למדוד ביטוי גנים לקבוצה גדולה יותר של חולים. להלן התוצאות:



ג. לכמה תתי קבוצות מתחלקים הפציינטים עכשיו?
ארבע קבוצות.

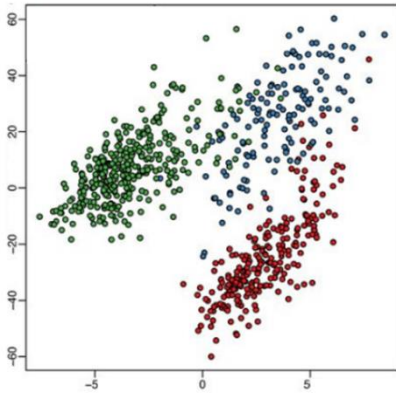
ד. על פי הנתונים הקיימים, מה המספר המינימלי של גנים שנדרשים על מנת לסווג פציינט חדש לאחת מהקבוצות הנ"ל ברמת בטחון סבירה?

ארבעה גנים. בכל קבוצת חולים כמעט כל הגנים מתבטאים גבוה באותה קבוצה ונמוך בשאר הקבוצות. לכן מדידת גן אחד המאפיין כל קבוצה תתן לנו אינדיקציה לאיזו קבוצה הוא שייך (נצפה לקבל 3 גנים עם ביטוי נמוך ואחד עם ביטוי גבוה).

ה. האם יש סיבה למדוד יותר מהמספר המינימלי של הגנים שהזכרו בשאלה הקודמת? הסבירו.
מכיוון שיש רעש במדידות ושונות בין אנשים כדאי לקחת יותר מהמספר המינימלי של הגנים.

שאלה פתוחה 4

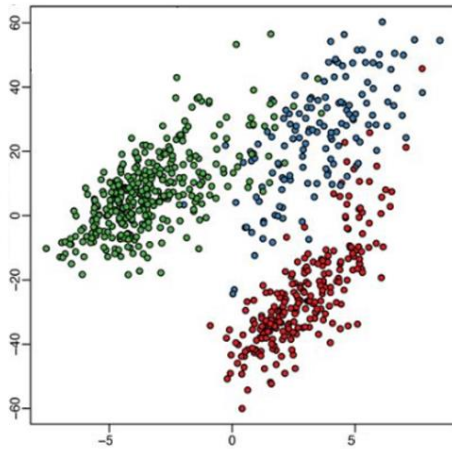
בגרף משורטטות 3 קבוצות על פי מדידות של 2 פרמטרים (אדום, ירוק, כחול).



- א. האם בעזרת k-means ומרחק אוקלידי הייתם מצליחים להפריד בין הקבוצות? הסבר איפה יתכן ויהיה קושי.
בסה"כ k-means יצליח להשיג הפרדה די טובה, הקבוצה הכחולה די מפוזרת ויש אזורים בהם ההפרדה בין הקבוצה הכחולה לשתיים האחרות לא תהיה טובה ודוגמאות יסווגו לא נכון.
- ב. כיצד צפויה להשתנות ההפרדה בין הקבוצות בעזרת k-means אם נעלה את המרחק האוקלידי בריבוע לפני ביצוע האשכול?

. העלאת המרחק בריבוע תביא לכך שמרחקים גדולים יעשו גדולים יותר ביחס למרחקים קטנים. מרכזי הכובד של הקבוצות יהיו צפופים יותר, והמרחק בין הקבוצות יגדל. אך זה יקשה על הפרדה בין הנקודות האדומות והירוקות שעכשיו יהיו קרובות יותר ללב הקבוצה הכחולה.

ג. שרטטו על גבי הגרף משמאל איפה יעבור PC1 ו- PC2 באנליזת PCA. סמנו את הצירים.



ד. איזה מה- PCs ששרטטתם בסעיף הקודם יכול לשמש להפרדה בין הקבוצות.
PC2.