**Breast Cancer Wisconsin (Diagnostic)**

**דו"ח מסכם בפרויקט התנסות במדעי הנתונים (858355201)**

**מגישים:**

**אוראל עטיה 208284968**

**אביתר אזולאי 314624438**

**איילת גיספאן** **207823170**

**מנחה:**

**ד"ר אלכסנדרה סימנובסקי**

**יולי 2025**

**תקציר:**

בפרויקט זה ניתחנו את מאגר הנתונים **Breast Cancer Wisconsin(diagnostic)** במטרה לזהות קבוצות חולים בעלי מאפיינים דומים, ללא שימוש בתווית האבחנה (ממאיר או שפיר).

מאגר הנתונים כולל 569 נבדקים, לכל נבדק נמדדו 30 תכונות נומריות, המחולקות לשלוש קטגוריות: **Worst,Mean ,Std** .

במהלך העבודה ביצענו:

* ניקוי ונורמליזציה של כלל הנתונים באמצעות .StandardScaler
* ניתוח הנתונים, הבנה שיש צורך להתמקד בקבוצת הWorst ובחירה ממוקדת בשלושה תכונות מקבוצה זו שנמצאו כמשמעותיות במיוחד:  
  **.Area, Compactness, Concave\_points**
* הרצה של אלגוריתמי קלאסטרינג:
* **K-Means**
* **Agglomerative**
* **DBSCAN**
* **Hierarchical Clustering**
* הפחתת ממדים לצורכי המחשה באמצעות **PCA** ו־**t-SNE**

התוצאות הראו כי קיימת הבחנה מבנית מובהקת גם ללא שימוש בתווית הדיאגנוזה. אלגוריתם K-Means הניב **ציון Silhouette של כ־0.59**, המעיד על הפרדה טובה יחסית בין הקבוצות.

**מבוא:**

מאגר הנתונים Breast Cancer Wisconsin (Diagnostic) פותח במטרה לספק תשתית מחקרית לאבחון אוטומטי של גידולים בשד. הוא מבוסס על תכונות שהופקו מתמונות מיקרוסקופיות של דגימות ביופסיה מסוג Fine Needle Aspiration (FNA), ומתאר מאפיינים מורפולוגיים של גרעיני תאים.

מאגר הנתונים כולל 569 נבדקים:

* 212 עם גידול ממאיר (Malignant) .
* 357 עם גידול שפיר (Benign).

כל דגימה כוללת 30 תכונות מספריות כגון: רדיוס גרעין התא, חספוס, היקף, מספר שקעים בגרעין וכו'.

כמו כן דגימה מתויגת כ"שפיר" (Benign) או "ממאיר" (Malignant).

**שאלת המחקר:**

**האם על בסיס 2-3 תכונות ניתן להצביע על הבדלים מובהקים בין דגימות הגידול, ולהבחין בקיומם של אשכולות מובחנים בדאטה ללא שימוש בתווית הדיאגנוזה.**

בפרויקט הנוכחי, מטרתנו הייתה לבדוק האם ניתן לבצע הבחנה בין גידולים ממאירים לשפירים ללא שימוש בתווית הסיווג, אלא באמצעות ניתוח אשכולות לא מפוקחים על בסיס תכונות נבחרות בלבד.

**סקר ספרותי:**

במהלך השנים בוצעו מחקרים רבים שבחנו את הקשר בין מאפייני גרעין התא לבין רמת הסיכון הממאיר של גידולים בשד, במיוחד באמצעות ניתוחים חישוביים על Breast Cancer Wisconsin Diagnostic Dataset. אחד הכיוונים המרכזיים שהודגשו בספרות הוא חיפוש אחר תכונות גאומטריות וביומורפולוגיות שמביאות להפרדה מובהקת בין גידולים ממאירים לשפירים – גם ללא אבחנה קלינית.

במהלך העבודה הסתמכנו על מספר מחקרים מדעיים שניתחו את הדאטה סט, והשתמשנו בשיטות של למידת מכונה כדי לבחון אילו תכונות הן המבדילות ביותר בין גידולים ממאירים לשפירים. מטרת הסקירה הייתה לתמוך בבחירה מושכלת של תכונות עבור שלב הקלאסטרינג הלא מונחה שביצענו בפרויקט.

המחקר הראשון, שפורסם ב־PubMed Central (PMC, 2020), ביצע ניתוח סטטיסטי מעמיק של 30 תכונות גרעיניות מתוך הדאטהסט, והציג השוואות בין פרמטרים כמו area\_worst, concave\_points\_worst, perimeter\_worst, ו־compactness\_worst. המאמר כלל גרפים מסוג Heatmap ו־Scatter Plot שהמחישו את רמת הקורלציה של כל תכונה עם האבחנה הקלינית. תכונות מקבוצת ה־worst נמצאו כבעלות שונות גבוהה יותר בין הקבוצות ולכן כאינפורמטיביות יותר לניתוח. ממצאיהם הצביעו על כך ש־area\_worst ו־concave\_points\_worst בפרט הם בין המדדים הבולטים ביותר להבדלה בין גידולים שפירים לממאירים.

המחקר השני, שפורסם ב־ScienceDirect (2024), התמקד בשיטות מתקדמות של למידת מכונה מבוססת Ensemble ובחירת תכונות (Feature Selection) לצורכי אבחון מוקדם של סרטן השד. החוקרים השתמשו באלגוריתמים כמו Random Forest, Gradient Boosting ו־XGBoost, ובחנו את השפעתן של כל אחת מהתכונות על ביצוע המודלים. גם כאן עלו תכונות ה־area\_worst, concave\_points\_worst ו־compactness\_worst כמשפיעות ביותר על תוצאת המודל. המאמר הצביע על כך שהשילוב בין תכונות פיזיות (כמו שטח הגרעין) לתכונות צורניות (כמו שקעים בגבול) יוצר תשתית חזקה למודל חיזוי מדויק – בין אם מונחה או לא מונחה.

כמובן הסתמכנו על עוד מס' מחקרים נוספים שיובאו בביליוגרפיה אך אלו היו המרכזיים ובהתבסס על ניתוחים שביצענו ועל מחקרים קודמים, החלטנו להתמקד בשלוש תכונות עיקריות: area\_worst, compactness\_worst ו־concave\_points\_worst. תכונות אלו נמצאו כתורמות במיוחד להבחנה בין גידולים ממאירים לשפירים – גם לפי מדדים סטטיסטיים כמו PCA ו־Silhouette Score, וגם לפי תצוגות גרפיות שביצענו (כמו גרפי t-SNE ותרשימי פיזור בתלת־מימד). בנוסף, מצאנו שהבחירה בתכונות האלו נתמכת במחקרים עדכניים, שחיזקו את החשיבות שלהן בהפרדה בין קבוצות החולים. השילוב בין ממצאים מהדאטה, תצפיות חזותיות ותמיכה מהספרות, ביסס את הביטחון שלנו בבחירה הזו, והיה חלק חשוב בבניית ההמשך של הפרויקט.

**שיטות עבודה:**

**ניתוח ראשוני של הנתונים (EDA)**

ניקוי נתונים: נבדקו ערכים חסרים וכפילויות- שניהם לא נמצאו לכן לא היה צורך בטיפול מיוחד בעניין זה.

הסרת מזהים וסיווגים: הסרנו את עמודות ID ו-Diagnosis, את הדיאגנוזה הורדנו כיוון שכל מטרת הפרויקט הוא קלאסטרינג שהוא כלי של למידה לא מפוקחת ולכן השימוש בדיאגנוזה אסור(בהמשך כן עשינו בה שימוש אך ורק לצורך בדיקת התוצאות שלנו). את עמודת הID הסרנו כיוון שהיא לא תורמת מידע חיוני עבור הבדיקות.

בחירת תכונות:

בחרנו להתמקד בתכונות מקבוצת ה־ **worst** מאחר שהן מייצגות את המאפיינים החריגים והקיצוניים ביותר של גרעין התא – כגון גודל מקסימלי, חוסר אחידות בגבולות ועומק השקעים. תכונות אלו מדגישות אזורים שבהם לרוב מופיעים סימנים מוקדמים לממאירות, בניגוד לתכונות ממוצעות שעלולות לטשטש חריגות נקודתיות.

הבחירה בגישה זו נשענת גם על תובנות קליניות, המצביעות על כך ששינויים סרטניים נוטים להופיע באזורים הקיצוניים של הרקמה. לכן, תכונות ה־ worst צפויות לספק מידע חד, מהימן ומשמעותי יותר לצורכי אשכול (clustering) ואבחון בלתי מונחה.

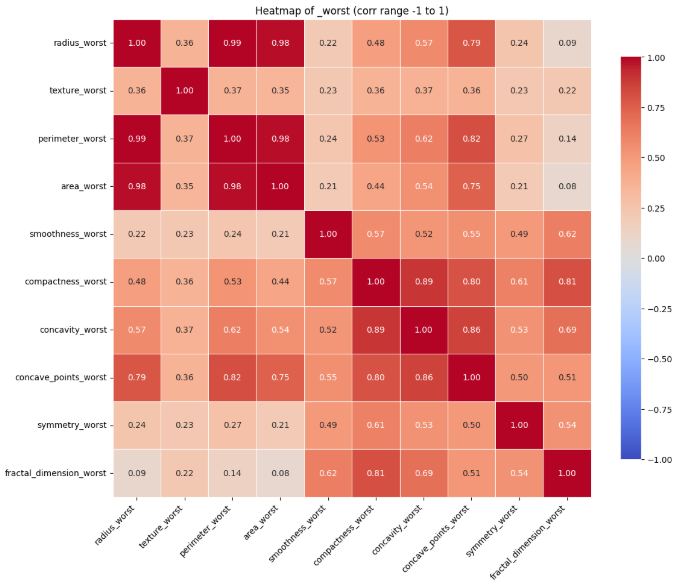
מקורות עבור הנחה זו:

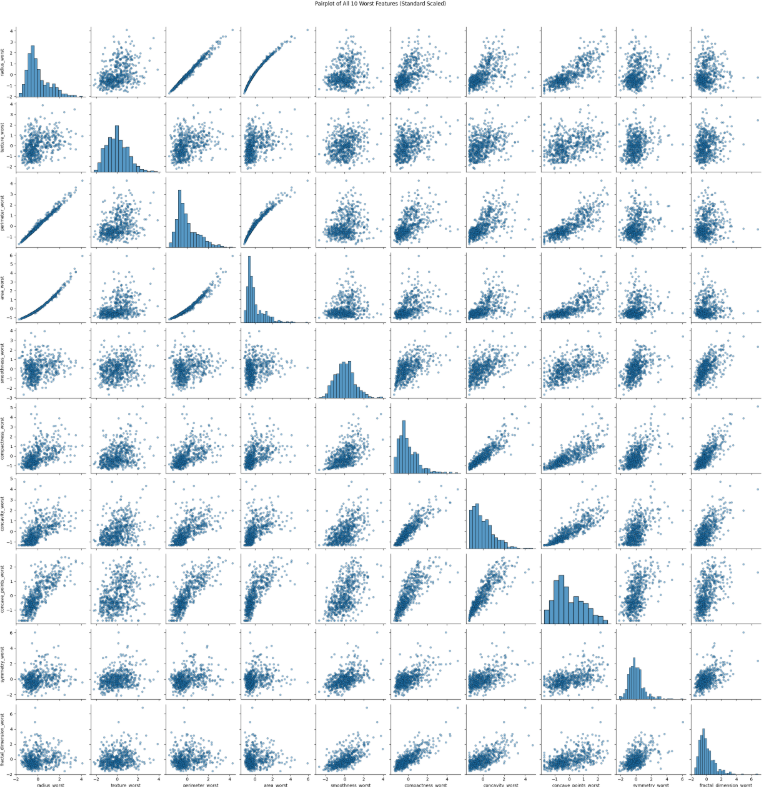
Pilapil ואחרים (2023) מצאו כי תאים בעלי מאפיינים מורפולוגיים קיצוניים נוטים להציג התנהגות אגרסיבית יותר, כולל פוטנציאל גבוה להתפשטות גרורתית. תוצאות אלו מדגישות את חשיבותם של מדדי קצה באבחון מוקדם של סרטן בעל מאפיינים מסוכנים.

* Brinkmann ואחרים (2023) הדגימו כי ערכים מקסימליים של תכונות כמו טקסטורה ועצימות בתמונת הגידול (texture & intensity) סיפקו דיוק גבוה יותר בסיווג גידולים לעומת מדדים ממוצעים. הממצאים שלהם תומכים בגישה שלפיה קצוות ההתנהגות של הגידול נושאים מידע קליני משמעותי יותר.

**בדיקת קשרים בין תכונות ה-Worst:**

כחלק מתהליך בחירת התכונות, יצרנו מפת חום (Heatmap) כדי לזהות מתאמים חזקים בין התכונות השונות. כאשר שתי תכונות מציגות מתאם גבוה, סביר שהן נושאות מידע דומה, ולכן ניתן להסיר אחת מהן ולשמור רק על התכונה היותר אינפורמטיבית – פעולה המסייעת בצמצום ממדים ומונעת כפילות בנתונים. בנוסף, השתמשנו בתרשים PairPlot כהשלמה חזותית למפת החום, במטרה לבחון לעומק את הקשרים בין זוגות תכונות. PairPlot מאפשר לנו לזהות דפוסי התפלגות, קשרים לינאריים בין משתנים, ואף רמזים למבנה אשכולות אפשרי – כל זאת באופן ויזואלי שתורם להבנה מעמיקה של הנתונים ולבחירה מושכלת של פיצ'רים להמשך הניתוח.



****

**הסקת מסקנות- Heatmap+Pairplot:**

בשלב בחירת התכונות, נעזרנו הן במפת חום והן בתרשים PairPlot כדי לזהות קשרים פנימיים בין משתנים ולהבין אילו תכונות עשויות להיות מיותרות או מועילות במיוחד. ממפת החום ניתן לראות מתאם חזק בין תכונות הקשורות לגודל הגידול – כמו radius, perimeter ו־area – ולכן בחרנו לשמור רק אחת מהן כדי למנוע כפילות. באופן דומה, זיהינו מתאם גבוה בין compactness, concavity ו־Concave\_points, ולכן שמרנו תכונה אחת מייצגת גם מקבוצה זו.

כך, על סמך מפת החום בלבד, ניתן היה לצמצם את מרחב התכונות באמצעות הסרת ארבע תכונות בעלות חפיפה גבוהה.

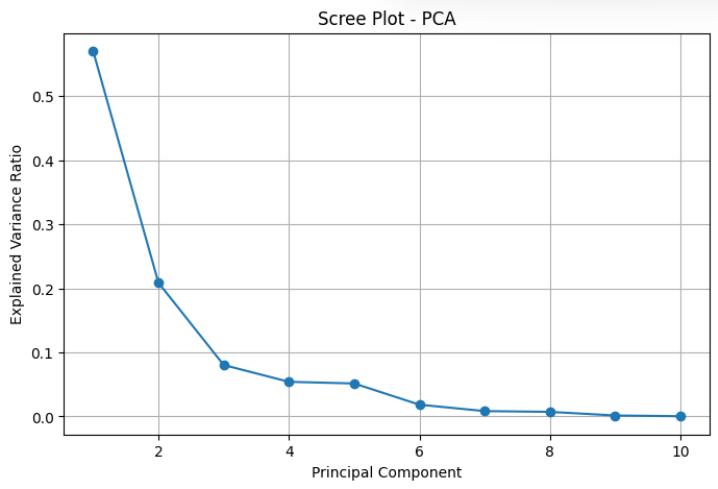
בנוסף, תרשים ה־PairPlot אפשר לנו לבחון באופן ויזואלי את הקשרים וההתפלגויות בין תכונות שונות, והוביל אותנו לבחור חמש תכונות שהראו שונות מובהקת והפרדה ברורה בין קבוצות פוטנציאליות:

* Area worst
* Concave points worst
* Fractal dimension worst
* Compactness worst
* Texture worst

תכונות אלו נבחרו הן בשל התרומה האינפורמטיבית שלהן והן בשל יכולתן האפשרית לשפר את אפקטיביות תהליך האשכול (clustering).

**PCA-הפחתת מימדים:**

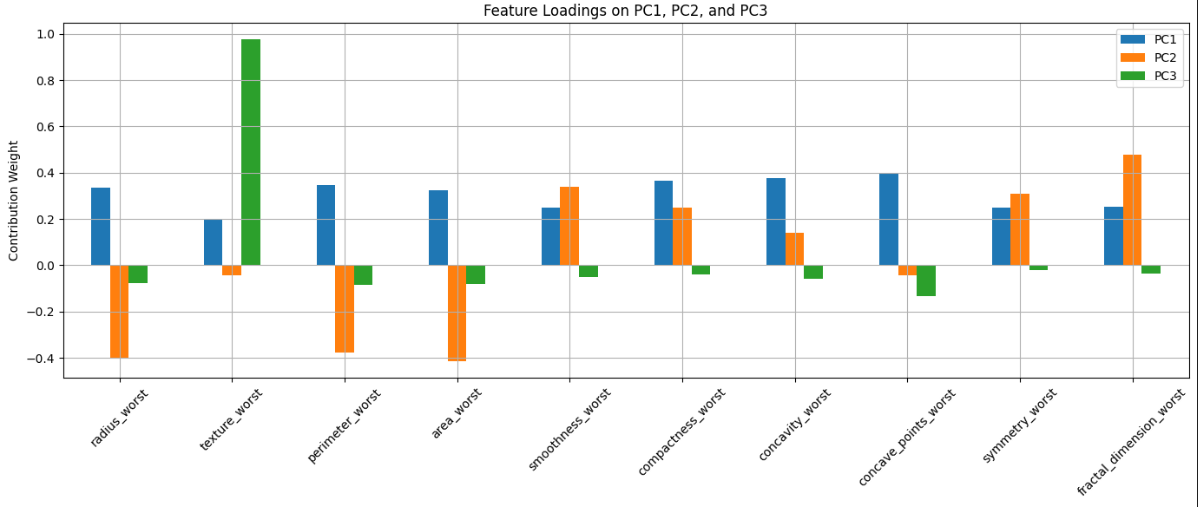
בשלב הבא, על מנת לפשט את מבנה הנתונים תוך שמירה על מידע משמעותי, יישמנו ניתוח רכיבים עיקריים (PCA). תחילה, בוצע נירמול לכל העמודות המספריות, כך שלכל תכונה יהיה ממוצע 0 וסטיית תקן 1 – פעולה המבטיחה שכל משתנה יתחשב באופן שווה בתהליך הפחתת הממדים.  
בהמשך, חישבנו את אחוז השונות שכל רכיב מסביר, והשתמשנו בגרף Elbow (ברך) כדי לקבוע בצורה חזותית את מספר הרכיבים האופטימלי לשימור, תוך איזון בין פשטות המודל לבין שמירה על מירב המידע שבנתונים.



מהגרף ניתן לראות כי 2 רכיבים עיקריים מסבירים 78% מהשונות ובמידה ונוסיף את השלישי נקבל הסבר לכ86% אחוז מהשונות.

לכן הסקנו כי מספיקים 2-3 תכונות בכדי לייצג את המידע המרכזי בדאטה בצורה יעילה, תוך הפחתת ממדים מבלי לאבד מידע מהותי.

כעת, נרצה לזהות את התכונות המשפיעות ביותר כלומר את אלו בעלות השונות הגדולה ביותר לאחר הורדת המימד, לכן נשתמש בגרף הבא- Feature loading:

****

בהתחלה חשבנו שהדרך הכי נכונה לבחור תכונות היא לקחת את המאפיין שהכי משפיע בכל אחד משלושת הרכיבים (PC1, PC2, PC3), כי כל רכיב מייצג כיוון אחר של שונות בתוך הנתונים. הרעיון היה שלקיחת תכונה אחת מכל רכיב תיתן לנו תמונה רחבה ומגוונת ככל האפשר.  
לפי זה, בחרנו את **texture**, **concave points** ,**fractal dimension**, שהיו עם ההשפעה (loading) הכי גבוהה בכל רכיב.

אבל כשבדקנו את התוצאה של הקלאסטרינג, ראינו שזה לא עבד טוב, המדד של  **silhouette** יצא נמוך, כלומר החלוקה בין קבוצות לא הייתה מוצלחת.  
בשלב הזה הבנו שהגישה הזו לא בהכרח מביאה את התוצאה הכי טובה, וחזרנו לבדוק את הגרפים וההשפעות של שאר המאפיינים.

ככל הנראה כיוון שהמרכיב הראשון( (PC1הוא בעל משקל משמעותית גבוה מהשניים האחרים, לכן בחרנו את Area ואת Concave points שכן הם בעלי ההשפעה הגבוהים ביותר במרכיב הראשון.  
לאחר מספר בדיקות בהם בדקנו איזה תכונה שלישית הכי כדאית קיבלנו שהצירוף של Compactnessאיתם מביא לחלוקה הכי יעילה בין הקבוצות. לכן לסיכום בחרנו עבור הקלאסטרינג בתכונות הבאות:

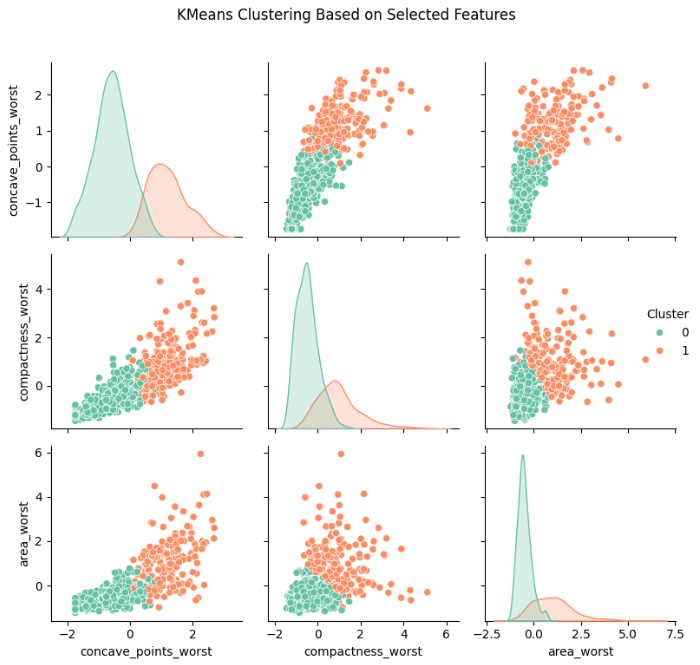
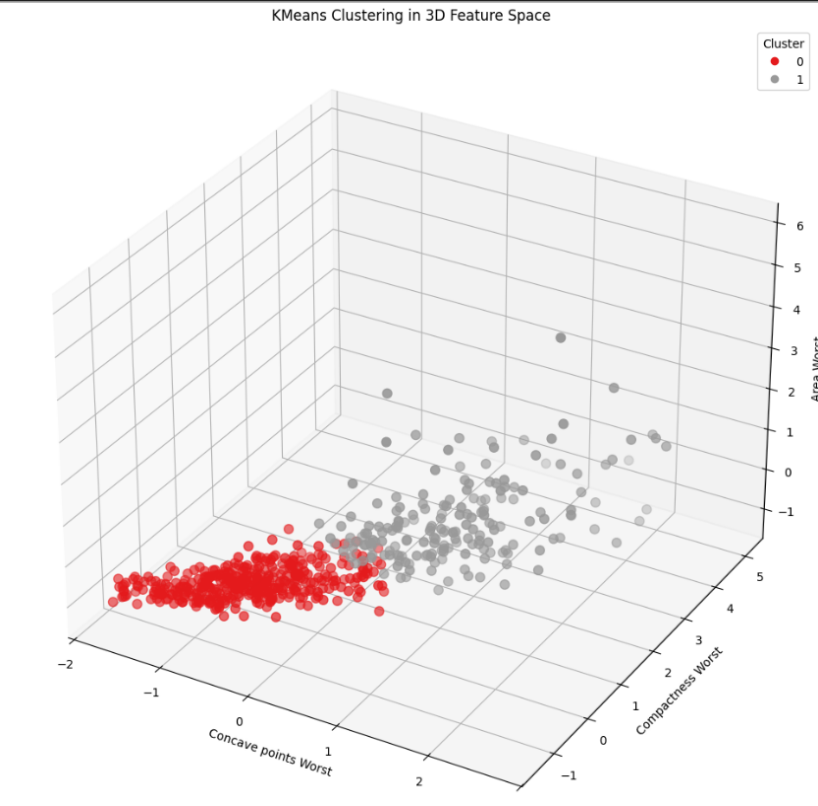
* **concave points**
* **Area**
* **Compactness**

**שלב הקלאסטרינג:**

בשלב זה החלטנו להריץ את האלגוריתמים עבור 2 קלאסטרים בלבד, ולא עבור 3 קלאסטרים. תחילה בחנו את הביצועים של כל אלגוריתם גם עבור חלוקה ל־2 וגם עבור חלוקה ל3 קבוצות, אך ניתוח ציון הSilhouette הראה כי ביצועי האלגוריתמים היו נמוכים יותר כאשר השתמשנו ב3 קלאסטרים. מאחר שציון הסילואט משקף את איכות ההפרדה והקומפקטיות של האשכולות, בחרנו להמשיך עם החלוקה ל2 קלאסטרים, שמספקת מבנה ברור ומוגדר יותר של הנתונים.

**K-Means:**

לאחר שביצענו הקטנת מימדים באמצעות PCA יישמנו את אלגוריתם K-Means כדי לחלק את הנתונים לקבוצות. מכיוון שהשיטה דורשת להגדיר מראש את מספר הקלאסטרים, נעזרנו בגרף Elbow כדי לאתר את המספר המתאים ביותר.  
הגרף הציג ירידה חדה בשונות בתוך הקבוצות ככל שמספר הקלאסטרים עלה, עד לנקודות K=2 ו־ K=3, שבהן נצפה שינוי משמעותי בקצב הירידה. מעבר לנקודות אלו, הירידה הייתה הדרגתית הרבה יותר, מה שמעיד שתוספת של קלאסטרים לא משפרת משמעותית את איכות החלוקה.

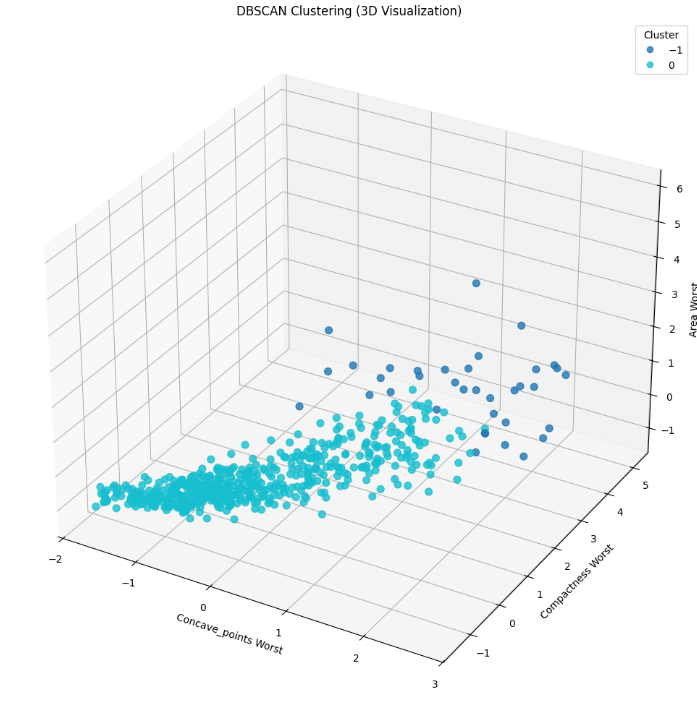
בשל העובדה שלא הייתה נקודה חד משמעית בגרף המרפק, החלטנו לבדוק את שתי האפשרויות 2 ו3 קבוצות על ידי הרצת האלגוריתם לכל ערך בנפרד. כדי להעריך איזו חלוקה מוצלחת יותר, השתמשנו במדד Silhouette אשר מודד את מידת ההתאמה של כל נקודה לקבוצה שלה ביחס לשאר הקבוצות. הרצנו את האלגוריתם הן עבור k=2 והן עבור k=3 וקיבלנו שעבור 3 קבוצות קיבלנו ערך יותר גבוה במדד, הערך שהתקבל היה 0.5369 שמעיד כי קיימת חלוקה די טובה בין הקבוצות כי שניתן לראות ויזואלית בגרפים להלן:

**:DBSCAN**

החלטנו להריץ גם אלגוריתם זה, למרות שבחרנו את הפרמטרים האופטימליים עבור אלגוריתם זה והתקבל ערך Silhouette גבוה יחסית (0.57), הוויזואליזציה מצביעה על תמונה שונה: מרבית הדגימות שובצו לקלאסטר יחיד, ומעט מאוד דגימות זוהו כרעש או כשייכות לקלאסטר נוסף. בפועל, לא נראית חלוקה ממשית בין קבוצות, והמבנה הקבוצתי אינו ברור לעין.

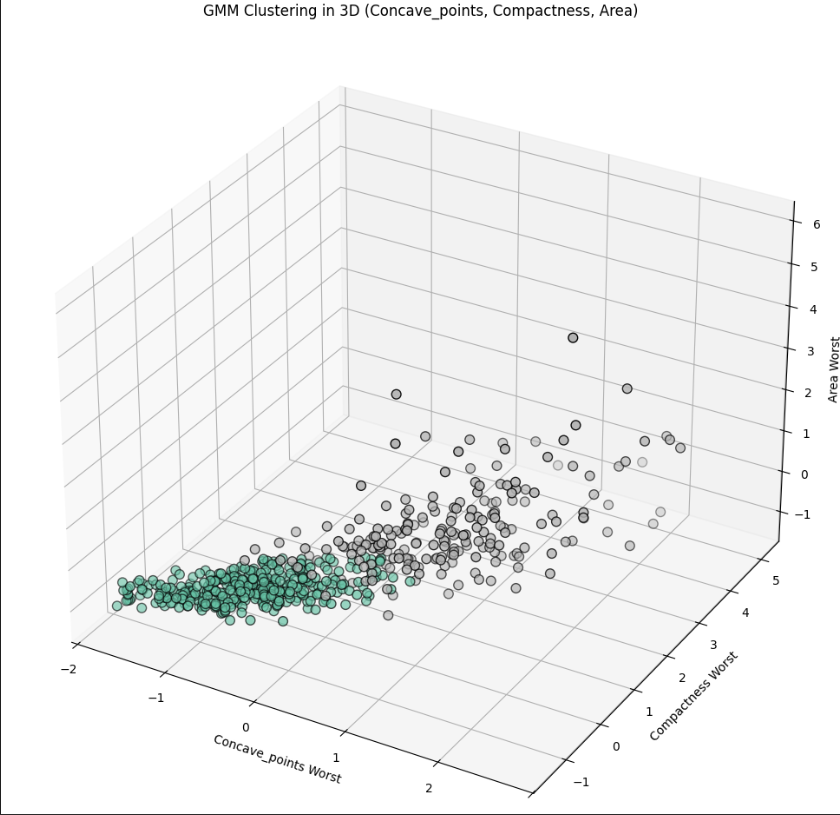
ממצאים אלה מצביעים על כך ש־ DBSCAN אמנם מזהה אזורים צפופים, אך עלול לאחד תתי מבנים לקבוצה אחת כאשר אין הפרדה ברורה ביניהם. ייתכן שגם לאחר הנרמול, הנתונים אינם מציגים גבולות חדים דיים שיאפשרו חלוקה ברורה בין קבוצות. ערך ה־ Silhouette הגבוה משקף את המרחק היחסי בין קבוצת הרוב לנקודות החריגות, אך לא בהכרח מעיד על חלוקה טבעית ומשמעותית.

מכך הסקנו שאין להסתמך על מדד ה־ Silhouette בלבד בעת הערכת איכות הקלאסטרינג, ויש לשלב תמיד גם ניתוח ויזואלי והבנה של מבנה הנתונים בעת בחינת תוצאות האלגוריתם.



**GMM(Gaussian Mixture Models):**

בחרנו להריץ את GMM, כי הוא מאפשר חלוקה גמישה יותר של הנתונים לעומת K-Means. במקום לחלק לפי מרחק בלבד, GMM בודק מה הסיכוי שכל נקודה שייכת לכל קבוצה וזה שימושי במיוחד כשאין גבול ברור בין הקבוצות. בדאטהסט של ייתכן שחלק מהתאים לא שייכים בבירור לקבוצה אחת (שפירה או ממאירה), ו־GMM יכול לעזור לזהות מקרים כאלה בצורה עדינה ומדויקת יותר, קיבלנו כי הערך של מדד Silhouetteהוא .0.502 ייתכן שהציון כאן מעט נמוך יותר בגלל האופן שבו הוא מחשב את ההשתייכות לקבוצות. בניגוד ל־K-Means שממקם כל נקודה ישירות בקבוצה אחת אלגוריתם זה מחשב הסתברות להשתייכות ולכן ייתכנו יותר נקודות גבוליות, שפוגעות בציון. כמו כן כיוון שהאלגו הנ"ל מניח התפלגות גאוסיאנית והנתונים שלנו הם נתונים רפואיים שיכולים להכיל שונות גבוהה, חריגות ותת-קבוצות במיוחד אצל תאים ממאירים, גורם שגם יכול להוריד את איכות החלוקה לפי המדד הזה.

נצרף כאן את הגרף ויזואלית: 

**:Agglomerative Clustering**

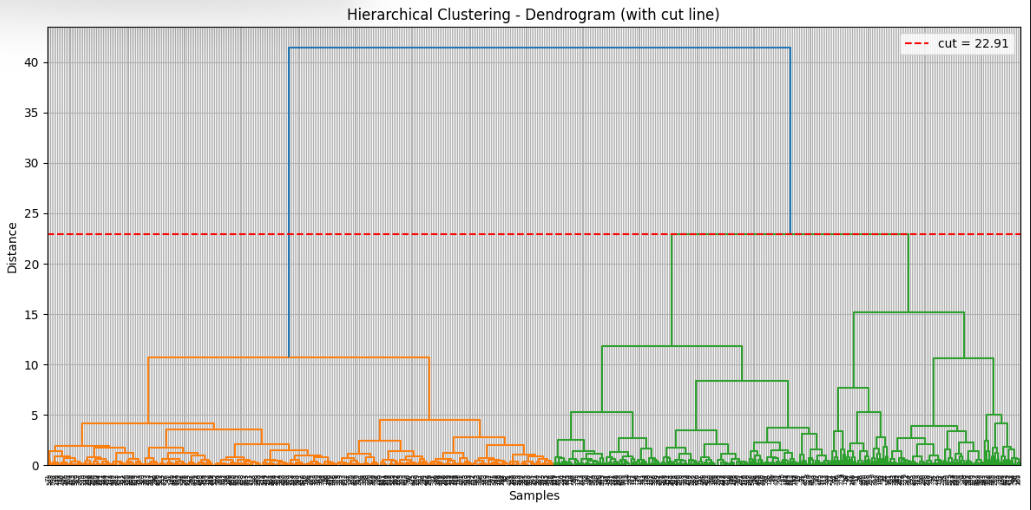
בחרנו לבדוק גם את האלגוריתם ההיררכי שמבצע חלוקה של הנתונים בצורה הדרגתית מהנקודות הקרובות ביותר ועד לאיחוד מלא של כל הדאטה.

האלגוריתם הזה מתאים לדאטהסט כמו שלנו, כי ייתכן שיש בו מבנים הדרגתיים או תת-קבוצות בתוך קבוצות (למשל, סוגים שונים של תאים ממאירים או שפירים). למרות היתרונות, קיבלנו שציון ה־Silhouette הוא 0.44 מעט נמוך יותר מאחר שהשיטה לא תמיד מחלקת את הנתונים בצורה הכי חדה, אלא לפי מבנה היררכי. לכן, נקודות שקרובות לשתי קבוצות שונות עלולות להישאר פחות ברורות, וזה משפיע על איכות הציון.

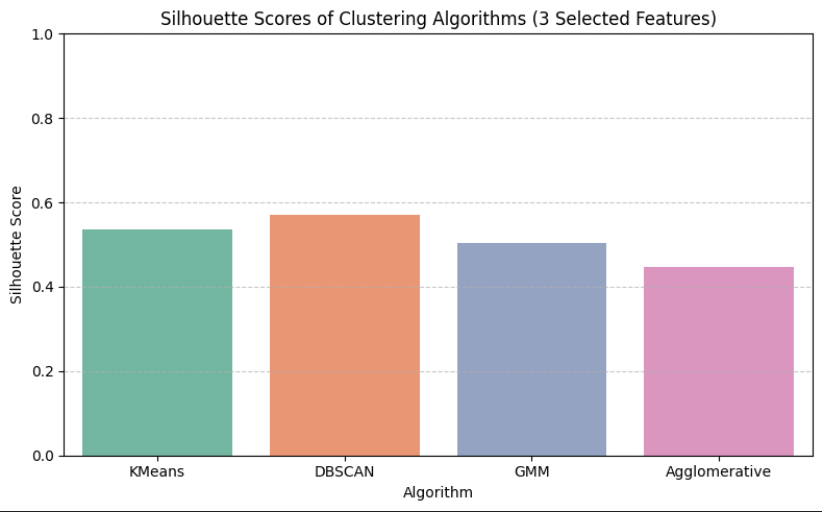
נסביר מעט על מה שקיבלנו: בתרשים הדנדוגרמה להלן ניתן לראות את קו החיתוך האדום שמסומן בגובה של כ־22.91. קו זה מייצג את הסף שבו בחרנו לחתוך את המבנה ההיררכי של העץ כלומר, הנקודה שבה אנו מחליטים לעצור את איחוד הקבוצות ולבחור את מספר הקלאסטרים הסופי.

מהדנדוגרמה ניתן לראות בבירור פיצול לשני ענפים עיקריים, דבר המחזק את ההנחה כי קיימות שתי קבוצות דומיננטיות בדאטה – תוצאה שמשרתת את ההנחה שלנו לאורך הדרך כי על בסיס התכונות שבחרנו ניתן לחלק את הדאטה ל2 קלאסטרים. כמו כן, ניתן להבחין בתתי אשכולות פנימיים בתוך כל קבוצה, דבר שעשוי לרמז על מבנים מורכבים יותר או תתי סוגים בתוך כל קטגוריה עיקרית של התאים, אך כפי שכתבנו לעיל הדבר העיקרי שבולט פה זה החלוקה לשני קבוצות עיקריות ובזה בחרנו להתמקד.

נצרף את הדנדוגרמה:



**סיכום תוצאות:**

****

יצרנו גרף המציג את ציוני ה־ Silhouette שהתקבלו עבור ארבעת האלגוריתמים השונים שאותם הרצנו על שלוש התכונות שבחרנו:

* \*DBSCAN: 0.57
* K-Means: 0.54
* GMM (Gaussian Mixture Model): 0.50
* Agglomerative Clustering: 0.44

**הסברת התוצאות:**

K-Means : הוא אלגוריתם שמחלק את הנתונים למספר קבוצות קבוע מראש, על סמך קרבה גיאומטרית למרכזים. ביישום על הדאטה שלנו, האלגוריתם סיפק חלוקה ברורה ונקייה לשני קלאסטרים עיקריים, עם ציון Silhouette גבוה יחסית של 0.54. הוויזואליזציה תמכה בהפרדה זו, והעובדה שכל התצפיות שובצו לקבוצות אפשרה לנו לראות את המבנה הכללי של הנתונים בצורה טובה. בגדול האלגוריתם תאם היטב להנחת קיומן של שתי קבוצות (ממאירים ושפירים).

DBSCAN : מתבסס על צפיפות נקודתית כדי לזהות קבוצות באופן דינמי ללא צורך לקבוע את מספר הקלאסטרים מראש.  
בדאטה שלנו, למרות שקיבלנו ציון Silhouette גבוה (0.57) החלוקה בין הקבוצות לא נראתה ברורה כשצפינו בגרף חלק גדול מהנקודות נשארו מבודדות או לא שויכו לקבוצה מסודרת. ייתכן שהסיבה לכך היא שאלגוריתם זה טוב בזיהוי קבוצות קטנות וצפופות, אבל פחות מצליח כשהגבולות בין הקבוצות מטושטשים או כשהמרחקים בין נקודות משתנים מאוד כמו בדאטה הרפואי שלנו.  
לכן, למרות שהמדד נראה גבוה, **בפועל** לא קיבלנו חלוקה שניתן להסיק ממנה מסקנות ברורות, ולא ניתן להסתמך על התוצאה הזו ככלי אמין לניתוח המבנה של סוגי התאים במקרה שלנו.

:Agglomerative Clusteringפועל בגישה היררכית, שבה כל נקודה מתחילה כקלאסטר עצמאי, והאלגוריתם מאחד בכל שלב את הזוג הקרוב ביותר עד להשלמת מבנה עץ מיזוג (דנדרוגרמה).  
בדאטה שלנו, התקבלה חלוקה ברורה לשתי קבוצות עיקריות, עם רמזים לתת קבוצות פנימיות בתוך כל אשכול תוצאה שמתחברת היטב לאופי המורכב של תאים שפירים וממאירים.  
למרות שציון הSilhouette היה נמוך יותר בהשוואה לאלגוריתמים אחרים, המבנה הוויזואלי שהתקבל נראה עקבי והגיוני, כאשר קו החיתוך בדנדרוגרמה מצביע על שני ענפים עיקריים.  
מכאן, ניתן להסיק כי האלגוריתם מצליח לחשוף מבנים עמוקים יותר בדאטה, גם אם המדדים המספריים לא מצביעים על איכות חלוקה מיטבית.

:GMM זהו אלגוריתם שמניח שהנתונים מגיעים מכמה התפלגויות גאוסיאניות וכל קבוצה מתוארת על ידי פרמטרים סטטיסטיים כמו ממוצע וסטיית תקן.  
כאשר הרצנו אותו על הדאטה שלנו, התקבלה חלוקה לשתי קבוצות עיקריות שנראתה הגיונית גם בגרף. ציון ה־Silhouette שקיבלנו היה 0.50 תוצאה טובה שמעידה על התאמה לא רעה למבנה הנתונים.  
יתרונו הבולט של האלגוריתם הוא בגישה ההסתברותית שלו, כל נקודה מקבלת סיכוי להשתייך לכל קבוצה, מה שמאפשר גמישות ודיוק גבוה יותר במיוחד במקרים גבוליים.  
לכן אלגוריתם זה הצליח לזהות בצורה יעילה את מבנה הקבוצות בדאטה שלנו ולהציע חלוקה רלוונטית.

**מסקנה:**

במהלך הפרויקט בדקנו את האפשרות לחלק את הדאטה למספר קבוצות שונות, ללא תלות בתווית האבחנתית. בתחילה לא ידענו מהו מספר הקבוצות המדויק, אך ביצענו תהליך EDA מסודר שכלל קריאת חומר רפואי והבנת המאפיינים שעשויים להבחין בין תאים ממאירים לשפירים. זיהינו כי תכונות מסוימות כמו concave points, area ו־compactness משפיעות באופן מובהק על מבנה הדאטה.

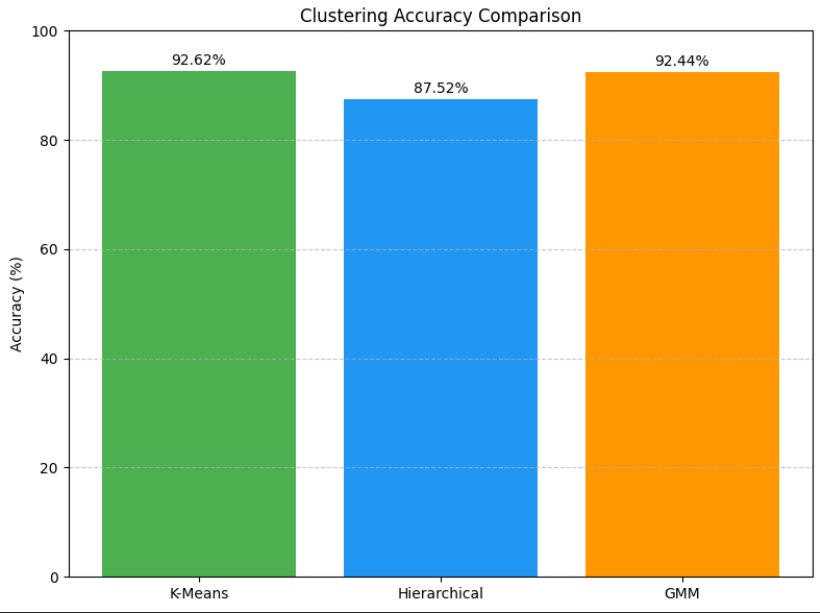
באמצעות כלים כמו Heatmap, Pairplot ו־PCA, זיהינו דפוסים ברורים בתצפיות. לאחר מכן הרצנו מספר אלגוריתמים של קלאסטרינג, אשר בעיקר שניים מתוכם שהם K-Mean ו־GMM הציגו חלוקה ברורה ויזואלית וקיבלו ציון גבוה יחסית במדד Silhouette.

הממצאים הצביעו על כך שאכן ניתן לזהות חלוקה משמעותית לשתי קבוצות עיקריות, באופן שתואם את ההנחה שלנו. השילוב בין הבנת הנושא וקריאת מידע בתחום הדאטהסט ויישום של אלגוריתמים מתאימים איפשר לנו להגיע לתובנה ברורה: אכן אפשר לזהות מבנה סמוי בנתונים גם ללא תוויות, ולהבדיל בין קבוצות בעלות משמעות קלינית.

**בדיקת נכונות:**

אמנם המשימה העיקרית בפרויקט הייתה לבצע Clustering ולהראות כי ניתן לזהות קבוצות נפרדות בתוך הדאטה, כפי שאכן הצלחנו להוכיח, אך החלטנו לקחת את הניתוח צעד אחד קדימה.  
לאורך כל הדרך הנחנו כי הקבוצות שהתקבלו מייצגות את החלוקה בין תאים ממאירים לשפירים, אך מאחר שקלסטרינג זוהי למידה בלתי מונחית לא יכולנו להשתמש בתווית הדיאגנוזה במהלך הלמידה ולכן ההנחה נותרה בגדר השערה בלבד.

עם זאת, מכיוון שתווית הדיאגנוזה כן הייתה זמינה לנו בדאטהסט, מצאנו לנכון להשוות בין הקבוצות שנוצרו לבין הערכים האמיתיים של סוג הגידול, כדי לבדוק האם אכן התקבלה הפרדה לפי ממאיר מול שפיר.  
לשם כך, השתמשנו במטריצת בלבול (Confusion Matrix) ככלי השוואה והערכה, ובאמצעותה קיבלנו את הנתונים הבאים:



הגרף שלפנינו מציג את אחוזי הדיוק שקיבלנו לאחר שהשווינו את תוצאות הקלאסטרינג שלנו לתוויות האמיתיות של הדיאגנוזה (שפיר מול ממאיר). ניתן לראות כי שלושת האלגוריתמים שבחרנו הצליחו לזהות את המבנה הדו קבוצתי בדאטה בצורה מרשימה:

* קיבלנו דיוק של 92.62% עבור K-Means
* 92.44% עבור GMM
* 87.52% עבור האלגוריתם ההיררכי

תוצאות אלו גבוהות במיוחד עבור משימה של למידה לא מונחית, שבה לא ניתנות מראש תוויות לתצפיות. שמחנו לראות כי אכן קיבלנו חיזוק מהותי להשערתנו כי קיימת חלוקה פנימית מובהקת בדאטה, ושניתן לזהות את ההבדל בין ממאיר לשפיר גם מבלי להשתמש בתוויות האבחנה.

**רפלקציה על תהליך העבודה – מה למדנו, איפה טעינו, ומה שיפרנו:**

**אתגרים בתחילת הדרך:**

החלק הכי קשה ומאתגר היה להבין מה זה קלאסטרינג לקח לנו זמן להבין את ההבדל בין למידה מונחית כמו קלסיפיקציה ובין למידה לא מונחית כמו קלאסטרינג במקרה שלנו. בעקבות הקושי הזה לקח לנו המון זמן לנסח שאלת מחקר שמתאימה לקלאסטרינג עבור הדאטהסט שלנו.

עבודה עם הנתונים ללא הדיאגנוזה חייבה אותנו לשנות את צורת החשיבה כיוון שלא היה לנו מדד נכונות מובהק לראות אם אנחנו בכיוון לכן נדרשנו להבין האם התוצאה שלנו הגיונית על בסיס פיזור הנתונים, מבנה הקבוצות, והקשר למאפיינים המקוריים של הנתונים.

אחד השלבים העיקריים והמורכבים היה בחירת התכונות, תחילה להבין בכמה תכונות נרצה להשתמש עבור הקליסטור וכמובן במי מהתכונות לבחור, הקושי היה לדלות מידע מתוך מחקרים רלוונטיים ומקורות מידע ולהבין אילו תכונות הן הכי רלוונטיות ויועילו באופן מקסימלי לחלוקה בין הקבוצות. ניסינו כמה שילובים והרצנו כמה פעמים את האלגוריתמים הן עבור מספר קבוצות שונה והן עבור קומבינציות שונות בין התכונות כמובן לאחר EDA מעמיק שממנו הגענו עם הנחות מוצא כמו שAREA וconcave points ככל הנראה יהיו יעילים מאוד. כמו כן השתמשנו הן בPCA עבור הורדת ממד והן במדד הסילואט כדי לקבל אינדיקציה לטיב החלוקה בין הקבוצות עבור הקומבינציות השונות.

**תהליך הלמידה וההתנסות:**

חלוקת עבודה ושיתוף פעולה מרחוק- בתחילת העבודה התקשינו לשתף פעולה מרחוק גם אוראל וגם אביתר היו הרבה במילואים לאורך התקופה האחרונה ובהתחלה כל אחד ניסה לעבוד במחברת Jupiter נפרדת, וכיוון שאנחנו 3 אנשים זה הוביל לחוסר תיאום ואי עקביות וכפילויות מיותרות בקוד. ברגע שהחלטנו לרכז את הכל במחברת אחת ולהשתמש ב־GitHub העבודה זרמה בצורה טובה ומדויקת יותר אך לקח לנו זמן להבין איך לנהל נכון גרסאות, איך למזג שינויים ואיך לעבוד בצורה מסונכרנת. בהמשך למדנו לנצל את הכלי הזה בצורה מיטבית .

בחירת כמות וסוג תכונות- כדי להבין את ההשפעה של בחירות שונות (למשל מספר רכיבי PCA או פרמטרים שונים), ביצענו הרצות מרובות והשוואות שיטתיות בין תרחישים. לא ניסינו למצוא "תשובה אחת נכונה", אלא לבדוק אילו תבניות מתגלות ומתי הקיבוץ הופך למשמעותי יותר.

שימוש באלגוריתמים- לאורך הדרך ניסינו מספר אלגוריתמים לקלסטרינג KMeans,Hierarchical ועוד. גילינו שחלקם פחות מתאימים למבנה הנתונים שלנו, מה שהוביל אותנו ללמוד איך כל אלגוריתם עובד ואיזה סוג נתונים מתאים לכל אלגוריתם ובעקבות כך לבצע התאמות. ניתחנו כל תוצאה לעומק ובחרנו באלגוריתמים המתאימים לסוג הדאטה שלנו ושמספקים חלוקה טובה לקבוצות.

**טעויות שמהן למדנו:**

* בתחילה כללנו בטעות משתנים קטגוריאליים בתוך תהליך PCA כלומר הרצנו אותו על כל הדאטה מה שהוביל לפגיעה באיכות התוצאה, לאחר כמה ניסיונות וניתוח חוזר הבנו את מקור הבעיה והסרנו את הID והדיאגנוזה של כל נבדק.
* לאחר ביצוע PCA ראינו כי 2 מרכיבים תופסים נפח של כ75% מהשונות וחשבנו שזה מספיק ולכן התחלנו לנסות להריץ אלגוריתמים רק עם 2 תכונות מרכזיות אך לא קיבלנו לא חלוקה ברורה ויזואלית ולא ציון טוב במדד הסילואט. רק לאחר שהוספנו רכיב שלישי, שנשא עוד 10% מהשונות, קיבלנו חלוקה טובה. בעצם המסקנה היא שהפחתת ממד וקבלה של 2 מרכיבים שמכילים את מרבית השונות לא בהכרח יובילו לחלוקה טובה כמו אצלנו וצריך לבדוק לגבי הוספת מרכיב או שניים נוספים.

* באלגוריתם DBSCAN קיבלנו מדד סילואט הכי גבוה, בדקנו את ציון המדד לפני שבדקנו ויזואלית איך החלוקה נראית וחשבנו כי רק על סמך המדד ניתן להסיק כי התבצעה חלוקה טובה לקבוצות, אך לאחר שהצגנו ויזואלית את הנתונים ראינו כי אין כלל חלוקה לקלסטרים והתקבל רק קלסטר אחד ועוד כמה נקודות אחרום שהם היו "רעש". בעצם למדנו כי לא ניתן להסתמך בקבלת ההחלטות רק על מדד אחד כמו סילואט אלא צריך לבחון גם ויזואלית וגם בעוד אלגוריתמים נוספים בכדי לגבש החלטה מהימנה.

**כלים ויכולות שפיתחנו:**

* העמקנו את ההיכרות שלנו עם ניתוח נתונים לא מונחה – כולל צמצום ממדים (PCA), קלאסטרינג מסוגים שונים, והסקת מסקנות מתוך פלטים גרפיים ומדדים.
* שיפרנו את יכולתנו הטכנית בשפת Python ובספריות כמו sklearn, pandas, seaborn ו־matplotlib.
* למדנו לעבוד עם GitHub אשר הועיל בשמירה על תיאום, חלוקת עבודה מסודרת, ומניעת דריסות או חוסר עקביות.
* חידדנו את היכולת לבחון ולפרש גרפים הן מבחינת מבנה ויזואלית והן לפי ערכים סטטיסטיים ולבחור את הדרך הנכונה להציג נתונים באופן ברור ואמין.
* למדנו להיות דינאמיים לאורך הפרויקט, היה שלב בתחילת הדרך שלקחנו את הפרויקט לכיוון של קלסיפיקציה ודי התקדמנו בו ונאלצנו למחוק הכל ולהתחיל מהתחלה, ועשינו את זה די בזריזות, למדנו את המשימה מחדש וביצענו אותה כמו שצריך לפי הבנתנו ובהנחיית אלכסנדרה**.**

**בביליוגרפיה:**

קישורים לרקע ספרותי:

<https://www.sciencedirect.com/org/science/article/pii/S1546221824008543>

<https://pubmed.ncbi.nlm.nih.gov/32357391/>

הנחת המוצא-להתמקד בערכי הWorst :

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10080822/>

: <https://www.mdpi.com/2072-6694/15/3/898>