1 ניתוח אשכולות

חיזוי וסיווג תפקוד בלוטת התריס של נבדק באמצעות ניתוח אשכולות (CLUSTER ANALYSIS).

ישנם שלושה סיווגים, תפקוד תקין (*healthy*), פעילות יתר (*hyperthyroid*), ותת פעילות (*hypothyroid*).

ניתוח אשכולות הוא כלי חשוב שמאפשר לזהות קבוצות של מטופלים עם תסמינים ומאפיינים דומים על סמך נתונים רפואיים. זה יכול לסייע ברפואה מונעת, אבחון ותכנון טיפול מותאם אישית.

לדוגמה, אלגוריתם K-means יכול לחלק נבדקים לקבוצות כמו תפקוד תקין (*healthy*), פעילות יתר (*hyperthyroid*), ותת פעילות (*hypothyroid*) בהתבסס על רמות הורמונים, תסמינים ומדדים ביוכימיים אחרים. הנבדקים יכולים לקבל טיפול מותאם או להיות מנוטרים לזיהוי מצבם המתדרדר. אלגוריתם DBSCAN יכול לזהות "אזורים צפופים" של מטופלים עם קונסטלציה דומה של סימפטומים ותוצאות בדיקות, וכך לאפשר גילוי של תתי-קבוצות ומחלה ייחודיות שלא היו ידועות קודם. המידע הזה עשוי להנחות מחקר רפואי עתידי ולהשפיע על הטיפול הניתן.

1.1 – מדדי איכות לאשכולות

בכדי להעריך את איכות האשכולות המתקבלים מתהליך ניתוח אשכולות, נעשה שימוש במגוון מדדי איכות. מדדים אלו מאפשרים לבחון את מידת ההתאמה והאפקטיביות של החלוקה שנוצרה. להלן מדדים מרכזיים:

- **הומוגניות**: מדד זה בוחן את מידת הדמיון בין העצמים בתוך כל אשכול. אשכול הומוגני יכיל עצמים שהם דומים זה לזה במאפיינים שלהם. ככל שהומוגניות האשכול גבוהה יותר, כך האשכול נחשב לאיכותי יותר. זהו מדד חשוב להערכת איכות החלוקה, מאחר שהוא מבטיח שכל אשכול מייצג קבוצה של עצמים עם מאפיינים משותפים חזקים.

- **שלמות:** שלמות מתייחסת למידת ההבדלים בין אשכולות שונים. חלוקה איכותית תיצור אשכולות שבהם יש דמיון גבוה בתוך כל אשכול ושוני ברור בין האשכולות. מדד זה מבטיח שהאשכולות אינם מכילים תתי-קבוצות דומות מדי, אלא שכל אחד מהם מייצג קבוצה מובחנת וברורה.

- **מגמתיות:** מדד זה בוחן האם האשכולות שנוצרו מצביעים על מגמות או תופעות בלתי-טריוויאליות בנתונים. מגמתיות מתייחסת לגילוי דפוסים משמעותיים ומעניינים שלא היו ברורים קודם לכן. חלוקה טובה תאפשר לחשוף תובנות חדשות על הנתונים, מה שיכול לסייע בקבלת החלטות ובהבנה עמוקה יותר של המבנים הקיימים במידע.

- **סקלביליות**: יכולתו של האלגוריתם להתמודד עם מערכי נתונים גדולים ועם נתונים רב-ממדיים היא מדד חשוב. אלגוריתם ניתוח אשכולות טוב צריך להיות יעיל ולספק תוצאות במהירות גם כאשר נפח הנתונים גדל בצורה משמעותית.

- **התמודדות עם נתונים רועשים**: ניתוח אשכולות יעיל צריך להיות חסין לרעש ולשגיאות בנתונים. מדד זה בוחן עד כמה האשכולות רגישים לערכים חריגים ועד כמה הם מושפעים מרעש בנתונים.

- **יכולת הסתגלות לנתונים חדשים**: המודל צריך להיות גמיש וניתן להתאמה כדי שיוכל להתמודד עם נתונים חדשים מבלי להצטרך לבנות מחדש את האשכולות מההתחלה. מדד זה חשוב להערכת יכולתו של האלגוריתם לשמור על ביצועים טובים לאורך זמן.

שימוש במדדי איכות מאפשר לבחון את האשכולות שנוצרו בצורה מעמיקה ולהבטיח שהם מייצגים בצורה מיטבית את המבנה והדפוסים של הנתונים.

1.2 – בחירת אלגוריתם לניתוח אשכולות

בהתלבטות בין שני אלגוריתמים לניתוח אשכולות – K-means ו-DBSCAN, הגענו להחלטה לבחור ב-K-Means על בסיס התאמתו למטרות הפרויקט וסט הנתונים הקיים. אלגוריתם K-means מחלק את הנתונים לקבוצות מוגדרות מראש על בסיס המרחק למרכזי האשכולות. לדוגמה, הוא יכול להפריד מטופלים ערכים תקינים של TSH ממטופלים עם רמות חריגות, ואז להמשיך ולחלק את הקבוצה השנייה לתתי-קבוצות של פעילות יתר (*hyperthyroid*), ותת פעילות (*hypothyroid*). יתרונו הוא הפשטות והיכולת להתאים את מספר האשכולות, אך הוא מוגבל לצורות גיאומטריות רגילות. בהקשר של סיווג לקבוצות כמו בשאלה, יתכן שיהיה קשה לדעת מראש אם יש רק 3 אשכולות או יותר. לכן, לאלגוריתם DBSCAN עשוי להיות יתרון כי הוא לא דורש קביעה זו מראש. עם זאת, אם ידוע שהאשכולות הם בצורה גיאומטרית רגילה, K-means עשוי להיות יעיל יותר. בנוסף, K-means פשוט יותר ליישום וחסכוני יותר מבחינת משאבי חישוב, מה שעשוי להיות שיקול חשוב כשמדובר במערכות קליניות. מומלץ גם לבחון טכניקות משולבות או אלגוריתמים אחרים כמו גם לבצע וידוא צולב לאמוד את ביצועי האלגוריתמים השונים במקרה הספציפי. להלן הלבטים והשיקולים בבחירת האלגוריתם.

K-Means

הרעיון המרכזי מאחורי סיווג באמצעות אלגוריתם K-Means הוא שהמשתמש קובע מראש את מספר האשכולות (K) שרוצים לחלק את הנבדקים אליהם. בדוגמה זו, טבעי שנרצה להשתמש באלגוריתם לפי 3 אשכולות. האלגוריתם ממקם באקראי K "מרכזי אשכולות" במרחב הנתונים הרב-ממדי (ממדים כמו רמות הורמונים, תסמינים וכו'). כל נבדק (נקודת נתונים) משויך לאשכול הקרוב ביותר ממרכז האשכול במונחי מרחק. לאחר השיוך, האלגוריתם מחשב מחדש את מיקום מרכזי האשכולות כך שיהיו במרכז הגיאומטרי של כל הנקודות המשויכות אליהם. שלבים אלו חוזרים על עצמם עד להשגת יציבות, כלומר עד שהנקודות כבר לא משנות שיוך בין האיטרציות. במילים אחרות, K-Means מנסה למצוא את החלוקה האופטימלית ל-K אשכולות, כך שהמרחק הכולל של כל הנקודות ממרכז האשכול שלהן יהיה מינימלי.

לדוגמה, אם נתחיל עם 3 מרכזי אשכולות, האלגוריתם ישייך נבדקים עם מאפיינים דומים לאותו אשכול, וינסה ליצור סיווג לפי קבוצה של תפקוד תקין (*healthy*), פעילות יתר (*hyperthyroid*), ותת פעילות (*hypothyroid*) על סמך המאפיינים.

היתרון בקביעה מראש של K הוא שהתוצאה תהיה תמיד מספר זה של אשכולות נפרדים. אך החיסרון הוא שאין כלי לזהות אם יש "אשכול רעש" של נקודות שאינן משתייכות לאף קבוצה מרכזית.

מאפיינים ותכונות

- פשטות ויעילות חישובית: אלגוריתם K-Means ידוע בפשטותו וביעילותו החישובית, מה שהופך אותו לאידיאלי למערכי נתונים גדולים.
- פרשנות ברורה: התוצאות של K-Means קלות להבנה ולפרשנות, מה שמאפשר זיהוי קל של קבוצות דומות במערך הנתונים.
- דרישות נתונים K-Means: דורש הגדרה מראש של מספר האשכולות, דבר המאפשר שליטה על מספר הקבוצות שהאלגוריתם יצור.

DBSCAN

הרעיון המרכזי מאחורי סיווג באמצעות DBSCAN הוא לאתר אזורים צפופים בנתונים הרפואיים של הנבדקים, כאשר כל אזור כזה מייצג אשכול או קבוצת מחלה. האלגוריתם סורק את מרחב הנתונים הרב-ממדי, כאשר כל ממד מייצג מאפיין רפואי כמו רמות הורמונים, תסמינים או תוצאות בדיקות אחרות. הוא מנסה לאתר אשכולות של נקודות נתונים צפופות מאוד, שמייצגות קבוצות של נבדקים עם מאפיינים דומים מאוד. שני הפרמטרים שהמשתמש קובע הם: רדיוס חיפוש מקסימלי ומספר נקודות נתונים מינימלי להגדרה כאשכול צפוף. כל קבוצה צפופה שנמצאה מוגדרת כאשכול נפרד, ונקודות בודדות מחוץ לאזורים הצפופים מסווגות כ"רעש".

לדוגמה, נניח שבמרחב הנתונים עם ממדים של רמות TSH, T4, חום וכדומה, נמצא אשכול צפוף של נבדקים עם רמות נורמליות של כל המדדים - זה יוגדר כאשכול "בריאים". לעומת זאת, אזור צפוף אחר של נבדקים עם רמות TSH נמוכות ותסמינים מסוימים, יסווג ככל הנראה כאשכול "*hypothyroid*".

המטרה היא שההחלטה על חלוקה לאשכולות תבוסס על צפיפות הנקודות במרחב הנתונים ולא על הנחות מוקדמות, מה שמאפשר לגלות דפוסים וקבוצות מחלה חדשים שלא היו ידועים מראש.

מאפיינים ותכונות

- התמודדות עם רעש וערכים חריגים: אלגוריתם DBSCAN מצטיין בזיהוי קבוצות צפופות תוך התעלמות מרעש וערכים חריגים, דבר שיכול להיות שימושי במערכי נתונים עם רעש.
- יכולת לזהות אשכולות בצורות לא רגילות/אי-גיאומטריות: DBSCAN מסוגל לזהות אשכולות בעלי צורות לא קונבנציונליות, מה שיכול להוות יתרון במערכי נתונים מורכבים.
- מזהה ומסווג נקודות "רעש" שאינן משתייכות לאף אשכול: בניגוד לאלגוריתמים אחרים *DBSCAN*, מאפשר להגדיר תצפיות שהן רחוקות יחסית מכל אותם אשכולות צפופים בתור "רעש". בהקשר שלנו, המשמעות של "רעש" תהיה נבדקים יוצאי דופן שאין להם שייכות באופן מובהק לשום אשכול. כך שסיווג האשכולות לא מושפע מנבדקים יוצאי דופן בצורה חריגה.

סיכום ביניים

הרעיון הכללי מאחורי שיטות לסיווג לאשכולות כמו K-Means ו-DBSCAN הוא לאתר דפוסים ומרכזי כובד בתוך מערך נתונים גדול, על בסיס דמיון של הנתונים ביניהם.

ב-K-Means, הגישה היא "מלמעלה למטה" - אנו מגדירים מראש את מספר האשכולות שרוצים למצוא, ואז האלגוריתם מחלק את הנתונים לאותם אשכולות על בסיס הדמיון/המרחק מנקודות המרכז שהוא ממקם. זוהי דרך לסווג לתוך קבוצות שהגדרנו מראש.

לעומת זאת, ב-DBSCAN הגישה היא "מלמטה למעלה" - האלגוריתם סורק את הנתונים ומנסה לאתר אזורים צפופים, שבהם הנתונים מתקבצים יחד בצורה טבעית. האשכולות מזוהות על בסיס האשכולות הצפופים הללו של נתונים דומים מאוד. זוהי דרך להרכיב אשכולות מתוך הנתונים עצמם, ללא הנחות מוקדמות.

בשני המקרים, המטרה היא להפריד את הנתונים לקבוצות הומוגניות של פריטים דומים על בסיס המאפיינים שלהם. אך ההבדל הוא שב-K-Means אנו מכתיבים את מספר הקבוצות מראש, ואילו ב-DBSCAN האלגוריתם מגלה באופן טבעי את הקבוצות המרכזיות מתוך הנתונים.

אם נחזור להקשר של סיווג נבדקים לפי מצב בריאותי, ניתן לראות שב-K-Means אנו מכריחים מראש חלוקה ל-3 קבוצות כי אלו הקבוצות שאנו מכירים. אבל ב-DBSCAN, האלגוריתם עשוי לגלות באופן טבעי יותר קבוצות או תתי-קבוצות שלא היינו מודעים אליהן קודם.

נימוק לבחירת K-Means

ההחלטה לבחור ב-K-Means נבעה מתהליך השוואה מדוקדק בין שני האלגוריתמים, בו נמצא כי אחוזי השגיאה של K-Means היו נמוכים באופן משמעותי בהשוואה ל-DBSCAN. תוצאות ההרצה של DBSCAN (עם הפרמטרים: $minPoints = 6$, $\epsilon = 0.9$) על סט הנתונים הראו כי האלגוריתם ייצר 189 אשכולות אך השאיר 2907 דוגמאות שלא סווגו, מה שהוביל לאחוז שגיאה של 49.08%. מנגד, K-Means סיפק תוצאות מדויקות יותר וחלוקה ברורה יותר של הנתונים לאשכולות, מה שהופך אותו לבחירה המועדפת לפרויקט זה.

ניתוח ופרשנות של תוצאות ההבדלים

התצפית על שיעור שגיאה גבוה יותר בשימוש באלגוריתם DBSCAN לעומת K-Means בהקשר הנתון מעלה מספר השערות אפשריות. ראשית, יש לתת את הדעת לטבען של הקבוצות במרחב הנתונים. DBSCAN מתבסס על זיהוי "אזורים צפופים" של נקודות דומות, ולכן הוא עשוי להתקשות אם האשכולות במקרה זה אינם מוגדרים היטב או בעלי צורה גיאומטרית לא רגילה. לעומת זאת, K-Means מכוון לאיתור אשכולות בצורות גיאומטריות רגילות יותר, מה שעשוי להתאים טוב יותר למבנה הנתונים הספציפי.

נקודה נוספת היא ש-DBSCAN מסווג נקודות יחידות שאינן משתייכות לאזורים הצפופים כ"רעש". אם יש שיעור גבוה של נקודות כאלה בנתונים, הדבר עלול להשפיע על דיוק הסיווג הכולל. K-Means, לעומת זאת, משייך כל נקודה לאחד האשכולות וככזה עשוי להציג שיעורי שגיאה נמוכים יותר במקרים מסוימים.

לבסוף, יש לקחת בחשבון גם את גודל מרחב הנתונים ומספר המאפיינים. DBSCAN עשוי להיות מורכב חישובית יותר ככל שמספר הממדים גדל, מה שעלול להשפיע על הדיוק.

פסאודו-קוד לאלגוריתם K-Means

1. קביעת מספר האשכולות (K) הרצוי על-ידי המשתמש.
2. בחירה אקראית של K "מרכזי אשכולות" ראשוניים במרחב הנתונים.
3. חישוב המרחק האוקלידי של כל נקודת נתונים מכל אחד מה-K מרכזי האשכולות.
4. שיוך כל נקודת נתונים לאשכול עם המרכז הקרוב ביותר אליה.
5. חישוב מחדש של מרכזי האשכולות על-ידי חישוב הממוצע של כל הנקודות המשויכות לאותו אשכול.
6. שלבים 3-5 אלו חוזרים על עצמם עד להשגת תנאי עצירה מוגדר מראש, לרוב כאשר אין עוד שינויים בשיוך הנקודות לאשכולות.

1.3 – תיאור שלבי ניתוח האשכולות עבור אלגוריתם K-Means

קובץ הנתונים לאחר עיבוד והכנה מכיל 9,172 רשומות בעלות 21 מאפיינים בנוסף למאפיין הנוסף, שהוא הערך המנובא (*Diagnosis*).

הכנת הנתונים

- **נורמליזציה:** נרמל את הנתונים כדי להבטיח שתכונות עם טווחים שונים לא ישפיעו באופן בלתי פרופורציונלי על התוצאות.
- **טיפול בערכים חסרים או שגויים:** נטפל בערכים חסרים או שגויים בנתונים (השלמה, המרה, הסרה) באותו האופן שעשינו בחלק הראשון של כריית המידע. יש לציין שבאלגוריתם K-Means ב-weka ערכים חסרים יוחלפו בערך ממוצע.
- **נרמול או סטנדרטיזציה של הנתונים:** נבצע נרמול וסטנדרטיזציה של הנתונים, כדי למנוע השפעה מוגזמת של משתנים בסדרי גודל שונים. נעשה זאת באותו האופן שעשינו בחלק הראשון של כריית המידע.

קביעת פרמטרים

- **מספר האשכולות (K):** יש לקבוע את מספר האשכולות הרצוי מראש. ננסה ערכים שונים של K ונבחר את לבחור את הערך המתאים ביותר. נריץ את האלגוריתם עבור ערכי k שונים, החל מ-3, מחמת שיש 3 קטגוריות לנבדקים, ועד ל-30. נמצא שהחלוקה הטובה ביותר היא עבור k=3.
- **פונקציית המרחק:** פונקציית המרחק באלגוריתם K-Means משמשת לחישוב המרחק בין נקודות הנתונים למרכזי האשכולות, דבר המאפשר קביעת ההשתייכות של כל נקודה לאשכול הקרוב ביותר. בהתבסס על הנתונים הקיימים, המדידה ע"י מרחק אוקלידי או מרחק מנהטן לא השפיעה על התוצאות בכלל.

- **קריטריון עצירה:** האלגוריתם ממשיך לרוץ עד להגעה לתנאי עצירה מוגדר, כמו מקסימום איטרציות או שינוי מזערי במיקום המרכזים. הגדרנו את קריטריון האיטרציות המקסימלי להיות 500.
- **אלגוריתם ליצירת המרכזים הראשוניים:** ישנם מספר שיטות לבחירת המרכזים הראשוניים של האשכולות באלגוריתם K-Means. הבחירה של שיטת ההתחלה משפיעה על ביצועי האלגוריתם ועל התוצאות הסופיות. להלן הסבר על השיטות השונות:
  - **בחירה רנדומלית (*Random Initialization*):** בשיטה זו, k נקודות במרחב הנתונים נבחרות באופן אקראי כמרכזי האשכולות ההתחלתיים. שיטה זו פשוטה אך עלולה להוביל לתוצאות לא אופטימליות אם הבחירה האקראית הייתה לא מייצגת. התוצאה הטובה ביותר שקיבלנו בשיטה זו הייתה עם רמת שגיאה של 25.93%.
  - **K-Means++:** זוהי שיטה מתוחכמת יותר לבחירת המרכזים ההתחלתיים. המרכז הראשון נבחר באקראי, ולאחר מכן הנקודות הבאות נבחרות כך שהסיכוי לבחירתן יהיה גדול יותר ככל שהמרחק שלהן מהמרכזים הקיימים גדול יותר. כך מובטח שהמרכזים מפוזרים היטב ברחבי הנתונים. שיטה זו משפרת במידה ניכרת את ביצועי האלגוריתם. התוצאה הטובה ביותר שקיבלנו בשיטה זו הייתה עם רמת שגיאה של 25.93%.
  - **Farthest First:** בשיטה זו, המרכז הראשון נבחר באקראי, והמרכז השני הוא הנקודה הרחוקה ביותר מהמרכז הראשון. המרכזים הנוספים נבחרים כך שתמיד הנקודה הרחוקה ביותר מכל המרכזים הקיימים היא שנבחרת כמרכז חדש. השיטה ממשיכה כך עד שכל ה-k מרכזים נבחרו. התוצאה הטובה ביותר שקיבלנו בשיטה זו הייתה עם רמת שגיאה של 40.14%.
  - **Canopy:** זוהי שיטת דגימה בשני שלבים. תחילה, מספר "חופות" (canopies) יוצרות קבוצות ראשוניות של נקודות על ידי הצמדת כל נקודה לחופה הקרובה אליה. לאחר מכן, K-Means רץ בנפרד על כל אחת מקבוצות החופות כדי לקבוע את המרכזים הסופיים של האשכולות בתוך כל חופה. שיטה זו יעילה במיוחד כאשר יש כמות גדולה של נתונים. התוצאה הטובה ביותר שקיבלנו בשיטה זו הייתה עם רמת שגיאה של 46.92%.

ההבדלים בין השיטות נעוצים ביעילות החישובית שלהן, ביכולת שלהן למפות את המבנה האמיתי של הנתונים ובהשפעה על התוצאות הסופיות של האשכולות. K-Means++ ובחירה רנדומלית הצליחו בגלל פיזור מיטבי של המרכזים הראשוניים, שהבטיחו התכנסות לפתרון כללי ויציב. לעומת זאת, Canopy ו-Farthest First ככל הנראה נכשלו בשל חוסר התאמה למבנה הנתונים הפשוט והברור, שהוביל לפיזור ראשוני לא יעיל של המרכזים הראשוניים.

ניתוח התוצאות

מהתוצאות עולה כי גישות K-Means++ ובחירה רנדומלית היו המוצלחת ביותר בניתוח האשכולות, עם שיעור שגיאה של כ-26%. סביר שהסיבה המרכזית לכך שאלגוריתם K-Means++ והבחירה הרנדומלית השיגו את התוצאות הטובות ביותר נעוצה ביכולת הייחודית של שיטות אלו להתמודד עם מערכי נתונים בעלי מבנה ברור ומוגדר, ובכך שהנתונים כנראה מסודרים בצורה כזו שההבדלים בין הקבוצות השונות בולטים ואינם מושפעים רבות מבחירת המרכזים הראשוניים. במילים אחרות, אם הקבוצות בנתונים הן נפרדות וברורות, גם בחירה אקראית של מרכזים יכולה להוביל להתכנסות מהירה ויעילה של האלגוריתם לפתרון מיטבי.

שיטת K-Means++ מתוחכמת יותר בכך שהיא מבטיחה שהמרכזים הראשוניים יהיו מפוזרים היטב, מה שמפחית את הסיכון להתכנסות למינימום מקומי ומשפר את יציבות האלגוריתם. אמנם המרכז הראשון נבחר באקראי, אך המרכזים הבאים נבחרים כך שהסיכוי לבחירתם גדל ככל שהמרחק שלהם מהמרכזים הקיימים גדול יותר. תהליך זה מאפשר פיזור אחיד של המרכזים הראשוניים ברחבי מרחב הנתונים, ובכך משפר את הסיכוי להתכנסות לפתרון כללי מיטבי. גישה זו מפחיתה את הסיכון להתכנסות למינימום מקומי ומשפרת את יציבות האלגוריתם, מה שהוביל לתוצאות מדויקות יותר. התוצאה הטובה שהתקבלה בשיטה זו משקפת את היכולת של K-Means++ לפזר את המרכזים הראשוניים בצורה אופטימלית, ובכך להוביל לחלוקה מדויקת של הנתונים.

לעומת זאת, הבחירה הרנדומלית, למרות פשטותה, הצליחה גם היא להתכנס לתוצאה זהה כמו במקרה של K-Means++. סביר להניח כי מבנה הנתונים במקרה זה היה יחסית ברור ואחיד, כך שגם בחירה אקראית של המרכזים הראשוניים הובילה לתוצאות טובות. כאשר הנתונים אינם רועשים, והקבוצות נפרדות זו מזו בצורה מובהקת, גם מרכזים שנבחרו באופן אקראי יכולים להוביל להתכנסות מהירה, חלוקה מדויקת, וסיווג יעיל של הנתונים.

סיכום ניתוח התוצאות

תוצאות אלו מעידות על כך שמבנה הנתונים היה נוח לסיווג, עם קבוצות מוגדרות היטב וגבולות ברורים בין הקבוצות. שתי השיטות הללו הצליחו להניב תוצאות טובות עקב התאמתן המיוחדת למבנה הנתונים הקיים. בעוד ש-K-Means++ מספקת פיזור ראשוני טוב יותר של המרכזים ומשפרת את יעילות האלגוריתם, הבחירה הרנדומלית הראתה שגם בפשטות ניתן להגיע לאותן התוצאות בתנאים מתאימים. עם זאת, חשוב לציין שבמערכי נתונים אחרים, ייתכן ושיטות אחרות יובילו לתוצאות טובות יותר, ולעיתים שילוב של מספר שיטות יכול להביא לתוצאות אופטימליות.

1.4 – דיווח תוצאות עבור אלגוריתם K-Means

סיכום התוצאות של הרצת אלגוריתם K-Means עבור סיווג נבדקים לקבוצות של תפקוד תקין (*healthy*), פעילות יתר (*hyperthyroid*), ותת פעילות (*hypothyroid*). ניתן לראות את הנקודות העיקריות הבאות:

מספר הנבדקים

האלגוריתם חילק את 9,172 הנבדקים לשלושה אשכולות.

מרכזי האשכולות הסופיים

- אשכול 0 (1,213 נבדקים): גיל ממוצע 52.75, מין נקבה, על תרופות תירוקסין, רמת TSH 5.85, רמת T3 1.97, רמת TT4 128.10, רמת T4U 0.999, רמת FTI 129.29.
- אשכול 1 (7,344 נבדקים): גיל ממוצע 52.31, מין נקבה, לא על תרופות תירוקסין, רמת *TSH 5.24*, רמת *T3 1.93*, רמת *TT4 104.45*, רמת *T4U 0.969*, רמת *FTI 110.31*.
- אשכול 2 (615 נבדקים): גיל ממוצע 48.30, מין נקבה, רמת *TSH 3.65*, רמת *T3 2.42*, רמת *TT4 121.16*, רמת *T4U 1.015*, רמת *FTI 122.60*.

```
Final cluster centroids:
                                          Cluster#
Attribute                     Full Data          0          1          2
                               (9172.0)   (1213.0)   (7344.0)    (615.0)
==============================================================================
age                              52.101    52.7461    52.3126    48.3008
sex                                   F          F          F          F
on_thyroxine                          f          t          f          f
query_on_thyroxine                    f          f          f          f
on_antithyroid_medication             f          f          f          f
sick                                  f          f          f          f
pregnant                              f          f          f          f
thyroid_surgery                       f          f          f          f
I131_treatment                        f          f          f          f
query_hypothyroid                     f          f          f          f
query_hyperthyroid                    f          f          f          t
lithium                               f          f          f          f
goitre                                f          f          f          f
tumor                                 f          f          f          f
hypopituitary                         f          f          f          f
psych                                 f          f          f          f
TSH                            5.2184     5.8504     5.2453     3.6503
T3                             1.9706     1.9672     1.9336     2.4201
TT4                          108.7003    128.101   104.4525   121.1597
T4U                            0.9761     0.9993     0.9689     1.0153
FTI                          113.6407   129.2913   110.3053   122.6026
```

מרכזי האשכולות הסופיים הראו דפוסים ברורים

- מאפייני תת פעילות בלוטת התריס (*hypothyroid*), גיל ממוצע 52.75, רמות גבוהות של *TSH* ו-*TT4*, כולל נבדקים על תרופות תירוקסין.
- אשכול 1: נבדקים בריאים, גיל ממוצע 52.31, רמות תקינות של רוב ההורמונים, לא על תרופות תירוקסין.
- אשכול 2: מאפייני יתר פעילות בלוטת התריס (*hyperthyroid*), גיל ממוצע 48.30, רמות גבוהות של *T3* ו-*FTI*, כולל נבדקים עם תכונות יתר פעילות בלוטת התריס.

השוואה לקבוצות המחלה האמיתיות

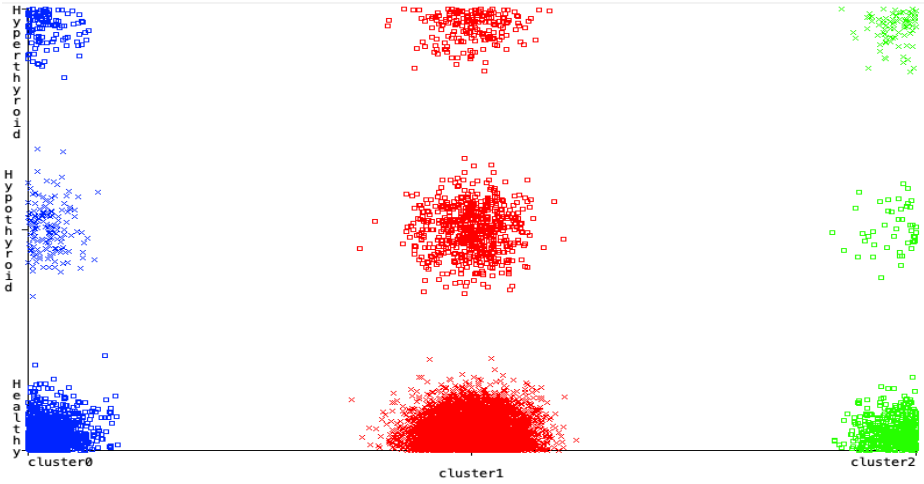
בהשוואה לקבוצות המחלה האמיתיות, 25.93% מהנבדקים סווגו באופן שגוי על ידי האלגוריתם.

התפלגות האשכולות

הסיווג הנכון לקבוצות היה:

- אשכול 0: 946 בריאים, 152 תת פעילות (*hypothyroid*), 115 פעילות יתר (*hyperthyroid*).
- אשכול 1: 6,559 בריאים, 598 תת פעילות (*hypothyroid*), 187 פעילות יתר (*hyperthyroid*).
- אשכול 2: 486 בריאים, 46 תת פעילות (*hypothyroid*), 83 פעילות יתר (*hyperthyroid*).

להלן התפלגות האשכולות לפי סיווג הנבדקים:



סיווג האשכולות

- אשכול 0: תת פעילות בלוטת התריס (Hypothyroid)
- אשכול 1: תפקוד תקין של בלוטת התריס (Healthy)
- אשכול 2: יתר פעילות בלוטת התריס (Hyperthyroid)

```
Cluster 0 <-- Hypothyroid
Cluster 1 <-- Healthy
Cluster 2 <-- Hyperthyroid
```

התפלגות הנבדקים לפי אשכולות

- נבדקים בריאים: מתוך 7,991 נבדקים בריאים, 6,559 סווגו לאשכול 1 (82.1%), 946 סווגו לאשכול 0 (11.8%), ו-486 סווגו לאשכול 2 (6.1%).
- נבדקים עם תת פעילות בלוטת התריס (*hypothyroid*): מתוך 796 נבדקים עם תת פעילות בלוטת התריס, 598 סווגו לאשכול 1 (75.1%), 152 סווגו לאשכול 0 (19.1%), ו-46 סווגו לאשכול 2 (5.8%).
- נבדקים עם יתר פעילות בלוטת התריס (*hyperthyroid*): מתוך 385 נבדקים עם יתר פעילות בלוטת התריס, 187 סווגו לאשכול 1 (48.6%), 115 סווגו לאשכול 0 (29.9%), ו-83 סווגו לאשכול 2 (21.5%).

```
Class attribute: diagnosis
Classes to Clusters:

    0    1    2  <-- assigned to cluster
  946 6559  486 | Healthy
  152  598   46 | Hypothyroid
  115  187   83 | Hyperthyroid
```

שיעור שגיאות הסיווג

מספר רשומות שלא סווגו כהלכה: 2,378 (25.93%)

1.5 – ניתוח תוצאות

האלגוריתם K-Means הצליח לחלק את הנתונים בצורה מסודרת לשלושה אשכולות, המבוססים על התכונות השונות של הנבדקים. חלוקה זו מהווה בסיס להבנת הדפוסים השונים של תפקוד בלוטת התריס ומספקת אינדיקציות ראשוניות לגבי יכולת הסיווג של האלגוריתם.

סקירת שיטת העבודה

אלגוריתם K-Means הוגדר עם שלושה אשכולות (K=3), תוך שימוש במרחק אוקלידי למדידת הקרבה בין הנקודות למרכזי האשכולות. כמות החזרות המקסימלית הוגדרה כ-500, ומרכזי האשכולות הראשוניים נבחרו על פי שיטת K-Means++. הנתונים כוללים משתנים קטגוריים ורציפים, עם החלפת ערכים חסרים בערכי הממוצע או השכיח.

סקירת התוצאות

מרכזי האשכולות מצביעים על ממוצעים של התכונות השונות בכל אשכול, ומסייעים בזיהוי הדפוסים המרכזיים.

- אשכול 0: הדפוסים מראים מאפיינים ברורים של תת פעילות בלוטת התריס עם גיל ממוצע של 52.75 ורמות גבוהות של TSH ו-TT4. הדבר מעיד על כך שהאלגוריתם הצליח לזהות קבוצת נבדקים עם תכונות תת פעילות בלוטת התריס.
- אשכול 1: נבדקים בריאים בגיל ממוצע של 52.31 עם רמות הורמונים תקינות. קבוצה זו היא הגדולה ביותר, מה שמעיד על התפלגות נתונים תקינה.
- אשכול 2: דפוסים של יתר פעילות בלוטת התריס, עם גיל ממוצע של 48.30 ורמות גבוהות של T3 ו-FTI, מראים כי האלגוריתם הצליח לזהות נבדקים עם תכונות יתר פעילות בלוטת התריס.

השוואה לקבוצות המחלה האמיתיות

שיעור השגיאות של 25.93% מצביע על כך שהאלגוריתם הצליח לסווג את רוב הנבדקים, אך יש מקום לשיפור. שיעור השגיאה הזה גבוה יחסית ומצביע על אתגר בסיווג המדויק של חלק מהנבדקים לקבוצות המחלה השונות יתכן וההנדסת תכונות נוספת ושימוש באלגוריתמים נוספים יכולים לשפר את הדיוק.

התפלגות האשכולות

בחינת התפלגות הנבדקים בכל אשכול לפי קבוצות המחלה האמיתיות מעלה מספר תובנות:

- אשכול 0: האשכול כולל בעיקר נבדקים עם תת-פעילות בלוטת התריס (152), אך גם 946 נבדקים בריאים ו-115 נבדקים עם יתר-פעילות. זהו אשכול לא הומוגני לחלוטין.
- אשכול 1: זהו האשכול הגדול ביותר, עם 6,559 נבדקים בריאים. עם זאת, ישנם גם 598 נבדקים עם תת-פעילות ו-187 עם יתר-פעילות שסווגו לאשכול זה.
- אשכול 2: באשכול זה ריכוז גבוה יחסית של נבדקים עם יתר-פעילות בלוטת התריס (83). עם זאת, ישנם גם 486 נבדקים בריאים ו-46 עם תת-פעילות שסווגו לאשכול זה.

לסיכום, רוב הנבדקים הבריאים סווגו נכונה באשכול 1, אך ישנה חפיפה מסוימת עם נבדקים מהקבוצות האחרות, ובנוסף, באשכול 0 קיימת התפלגות לא הומוגנית של נבדקים. הדבר מצביע על מורכבות בתכונות שיכולות להוביל לשגיאות סיווג.

סיווג האשכולות

חלוקה זו עוזרת להמחיש את היכולת של האלגוריתם לזהות תכונות עיקריות המשפיעות על תפקוד בלוטת התריס. האשכולות מצביעים על חלוקה ברורה, אך יש לשפר את הדיוק כדי להפחית את שיעור השגיאות.

התפלגות הנבדקים לפי אשכולות

ניתוח ההתפלגות מראה כי רוב הנבדקים הבריאים סווגו נכונה, אך ישנה חפיפה בין תכונות של נבדקים עם תת פעילות ויתר פעילות בלוטת התריס, מה שמעיד על חפיפה בתכונות ההורמונליות.

מסקנות

האלגוריתם הצליח לזהות את שלוש קבוצות המחלה העיקריות ואת המאפיינים ההורמונליים והגילאיים המרכזיים של כל קבוצה. עם זאת, שיעור השגיאה של כ-26% בסיווג הנבדקים לקבוצות מרמז על אתגר בהפרדה המדויקת של חלק מהנבדקים על בסיס המאפיינים שנבחנו. כל אחד מהאשכולות כולל תערובת של נבדקים משלוש קבוצות המחלה, אם כי עם דגש על קבוצה אחת מרכזית. ייתכן שקיימים גורמים או תתי-קבוצות נוספים שלא זוהו במודל הנוכחי. לשיפור התוצאות, ניתן לשקול שילוב של טכניקות נוספות כמו הפחתת ממדים, שימוש באלגוריתמים משלימים או הוספת משתנים ומאפיינים רלוונטיים נוספים. חשוב לבצע הערכה מקיפה של האלגוריתם באמצעות טכניקות כמו וידוא צולב כדי לקבל הערכה מדויקת יותר של הדיוק והיציבות. ממצאי המחקר יכולים לשמש בסיס לבחינה נוספת של גורמים גנטיים, סביבתיים ורפואיים העשויים להשפיע על התפתחות מצבי תירואיד שונים ועל הקשר ביניהם.

סיכום

סיכומו של דבר, ניתוח האשכולות סיפק תובנות ראשוניות חשובות לגבי מאפייני הנבדקים בקבוצות המחלה השונות. שיעור השגיאות מצביע על הצורך בשיפור הדיוק באמצעות הנדסת תכונות נוספת ושילוב של אלגוריתמים נוספים. כמו כן, ניתוח ההתפלגות מראה את חשיבות הגיל והמאפיינים ההורמונליים להבנת תפקוד בלוטת התריס ולסיווג נכון של הנבדקים. עם זאת, נדרש מחקר נוסף כדי לשפר את דיוק הסיווג, לזהות גורמי השפעה נוספים ולהעמיק את ההבנה של המנגנונים המניעים את התפתחות בעיות התפקוד השונות של בלוטת התריס.

ניתוח ארכיטקטורה וביצועים של רשת נוירונים מלאכותית לסיווג בעיות בלוטת התריס

נבחן את היישום של רשת נוירונים מלאכותית (Artificial Neural Network) לצורך סיווג בעיות רפואיות הקשורות לתפקוד בלוטת התריס. נתמקד בניתוח מבנה הרשת, תהליך האופטימיזציה, והערכת ביצועי המודל. המטרה היא להעריך את יעילות השיטה ולזהות תחומים פוטנציאליים לשיפור.

2.1 – הגדרת ארכיטקטורת הרשת

הרשת העצבית שנבחרה למטלת סיווג זו היא רשת Multilayer Perceptron (MLP), השייכת למשפחת רשתות הזנה קדמית (Feed-Forward Neural Networks). ארכיטקטורה זו נבחרה בשל יכולתה לייצג קשרים לא לינאריים מורכבים בין המאפיינים לבין הקטגוריות המוגדרות, מה שנדרש במטלות סיווג רפואיות מסוג זה.

מבנה הרשת

- שכבת קלט עם 21 נוירונים, אחד עבור כל מאפיין
- שתי שכבות נסתרות עם 10 נוירונים בכל שכבה (ניתן לשנות את מספר השכבות והנוירונים לאחר ניסיונות ראשוניים)
- שכבת פלט עם 3 נוירונים, אחד עבור כל קטגוריה (*hypothyroid*, *hyperthyroid*, *healthy*)

הרשת מורכבת משכבת קלט בעלת 21 נוירונים, המייצגים את 21 המאפיינים של כל דוגמה בקובץ הנתונים. לאחר שכבת הקלט, מצויות שתי שכבות נסתרות (Hidden Layers) עם 10 נוירונים בכל אחת מהן. שכבות נסתרות אלו מאפשרות למידה של ייצוגים פנימיים מורכבים של הקשרים בין המאפיינים לקטגוריות. לבסוף, שכבת הפלט מכילה 3 נוירונים, המייצגים את שלוש הקטגוריות האפשריות: *hypothyroid*, *hyperthyroid* ו-*healthy*.

הנוירונים בכל שכבה מחוברים באופן מלא (Fully Connected) לכל הנוירונים בשכבה העוקבת. כלומר, כל נוירון בשכבה מסוימת מקבל קלט ממשקלי החיבורים מכל הנוירונים בשכבה הקודמת. המידע יזרום מכיוון שכבת הקלט אל שכבת הפלט דרך השכבות הנסתרות.

פונקציית ההפעלה

פונקציית ההפעלה לנוירונים בשכבת הפלט תהיה פונקציית ההפעלה Sigmoid. המתוארת על ידי הנוסחה:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}} = 1 - \sigma(-x)$$

פונקציית ה-Sigmoid מחזירה ערכים בטווח שבין 0 ל-1, מה שהופך אותה למתאימה במיוחד למשימות סיווג בהן התוצאה מבטאת הסתברות.

יתרונות פונקציית ההפעלה Sigmoid במחקר זה:

- המרת ערכים להסתברויות: בפונקציית ה-Sigmoid, כל ערך קלט מומר לערך בטווח של 0 ל-1, המסמל הסתברות. בתהליך הסיווג, זה מאפשר לפרש את הפלט כהסתברות להיות שייך לכל אחת מהקטגוריות.
- רציפות וגזירות: פונקציית ה-Sigmoid רציפה וגזירה בכל התחום, דבר המאפשר לחישובי השיפוע (Gradient Descent) לעדכן את המשקלים בצורה יעילה במהלך תהליך הלמידה.
- טיפול בערכים קיצוניים: פונקציית ה-Sigmoid מגבילה את הפלט לערכים שבין 0 ל-1, כך שערכי קלט קיצוניים לא ישפיעו בצורה בלתי פרופורציונלית על התוצאה הסופית.

למרות יתרונות אלו, ישנם חסרונות פוטנציאליים לפונקציית ההפעלה Sigmoid, כמו קצב התכנסות איטי יותר בהשוואה לפונקציות אחרות (למשל ReLU) ובעיית ה-Vanishing Gradients, שעלול להאט את תהליך האימון. אולם, בהקשר של רשת ה-MLP שבנינו ולסוג הנתונים הרפואיים בהם אנו עוסקים, פונקציית ה-Sigmoid מספקת את האיזון הנדרש בין יציבות פרשנית ויכולת סיווג מדויקת.

2.2 – אופטימיזציה

פרמטרים

על מנת להשיג תוצאות אופטימליות בתהליך הלמידה של רשת הנוירונים, יש להגדיר בקפידה את הפרמטרים המרכזיים המשפיעים על האופטימיזציה. נתמקד בשלושה פרמטרים עיקריים: פונקציית השגיאה (Loss Function), גודל ה-Batch, וקצב הלמידה (Learning Rate).

פונקציית השגיאה (Loss Function)

בעת אימון הרשת, המטרה היא לצמצם את השגיאה של הרשת ככל הניתן על מנת שהחיזוי של הרשת יהיה קרוב יותר לתוצאות במציאות. הדרך להשפיע על גודל שגיאת האימון טמונה בשינוי המשקולות ובערכי ההטיה של הרשת כך שהרשת תפיק תוצאות חיזוי קרובות יותר לתוצאות האמת. פונקציית השגיאה משמשת למדידת הפער בין חיזוי וסיווג התוצאות שנוצרו על ידי הרשת לבין התוצאות האמיתיות. פונקציית השגיאה שנבחרה עבור רשת זו היא MSE (Mean Squared Error) הסטנדרטית.

MSE מחשבת את ממוצע ריבועי ההפרשים בין התוויות החזויות לתוויות האמיתיות. הנוסחה של MSE נתונה על ידי:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

כאשר n הוא מספר הדגימות, y_i היא התווית האמיתית של הדגימה ה-i, ו-$\hat{y}_i$ היא התווית החזויה על ידי הרשת עבור הדגימה ה-i.

מזעור פונקציית MSE תגרום לכך שחיזויי המודל יתקרבו ככל האפשר לערכי התוויות האמיתיות, ובכך תשפר את יכולת הסיווג הכוללת של הרשת. למרות שבמשימות סיווג רב-מחלקתיות, פונקציות שגיאה אחרות כמו Cross-Entropy עשויות להיות מתאימות יותר, נאלץ להסתמך על MSE בשל מגבלות המימוש של Weka.

גודל ה-Batch

גודל ה-Batch מגדיר את מספר הדגימות שהרשת "רואה" בכל שלב של תהליך הלמידה לפני שמתבצע עדכון של הפרמטרים של הרשת (המשקלים).

בחרנו בגודל Batch של 128, כלומר בכל איטרציה הרשת מעבדת 128 דגימות במקביל, מחשבת את השגיאה הממוצעת על פני דגימות אלו ומבצעת עדכון בהתאם. גודל ה-Batch משפיע על מהירות ויציבות תהליך הלמידה. Batch גדול מדי עלול להאט את התהליך ולהקשות על התכנסות לפתרון אופטימלי, בעוד Batch קטן מדי עלול להוביל לעדכונים תכופים ולא יציבים של פרמטרי הרשת. גודל של 128 נמצא כפשרה טובה המאזנת בין שני השיקולים הללו, בהתבסס על ניסוי וטעייה.

קצב הלמידה (Learning Rate)

קצב הלמידה הוא פרמטר קריטי בתהליך האופטימיזציה, המגדיר את גודל הצעד בכל עדכון של משקלי הרשת בכיוון השיפוע השלילי של פונקציית השגיאה. קצב למידה גבוה מדי עלול לגרום לאי-יציבות ולחוסר התכנסות, בעוד שקצב נמוך מדי יוביל לתהליך למידה איטי מאוד.

2.3 – הפעלת הרשת והערכת ביצועים

לאחר הגדרת ארכיטקטורת הרשת ופרמטרי האופטימיזציה, נתמקד כעת בהיבט המעשי של אימון הרשת והערכת ביצועיה. תהליך זה כולל ביצוע אימונים חוזרים (Epochs) על מערך הנתונים, תוך מעקב אחר מדדי הביצוע על קבוצת האימון וקבוצת המבחן (Validation Set).

חלוקת הנתונים

ראשית, נשים לב כי בקבצי מערך הנתונים יש שתי קבוצות נפרדות - קבוצת אימון (Training Set) וקבוצת מבחן (Test Set). חלוקה זו תאפשר לנו לאמן את הרשת על קבוצה אחת של נתונים, ולהעריך את ביצועיה על קבוצה בלתי תלויה שלא נחשפה במהלך הלמידה.

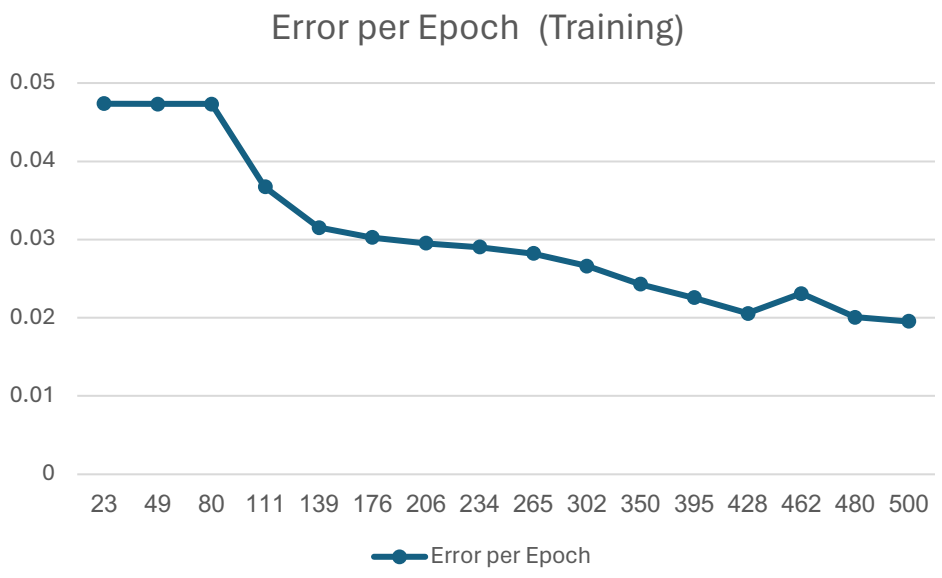
תהליך האימון

- נטען את קובצי הנתונים המוכנים מראש של קבוצת האימון וקבוצת המבחן.
- נגדיר את מספר ה-Epochs ל-500, בהתאם לערך שצוין בפרמטרים של תהליך האופטימיזציה.
- נאתחל את הרשת העצבית על פי הארכיטקטורה שנקבעה - 21 נוירונים בשכבת הקלט, שתי שכבות נסתרות עם 10 נוירונים כל אחת, ושכבת פלט עם 3 נוירונים.
- נבצע את תהליך האימון למשך 500 איטרציות (Epochs), כאשר בכל איטרציה הרשת תעבור על כל דגימות האימון, תחשב את פונקציית השגיאה (MSE), ותבצע עדכון של המשקלים באמצעות אלגוריתם האופטימיזציה.
- במהלך האימון, נחשב ונאגור את ערכי פונקציית המחיר (MSE) עבור קבוצת האימון וקבוצת המבחן בסיום כל Epoch.
- בנוסף, נחשב את מדדי הביצוע הבאים עבור שתי הקבוצות בסיום כל Epoch:
- אחוז הדיוק (Accuracy) - שיעור הדגימות שסווגו נכונה מתוך סך הדגימות.

גרף השגיאות (MSE) כפונקציה של Epochs

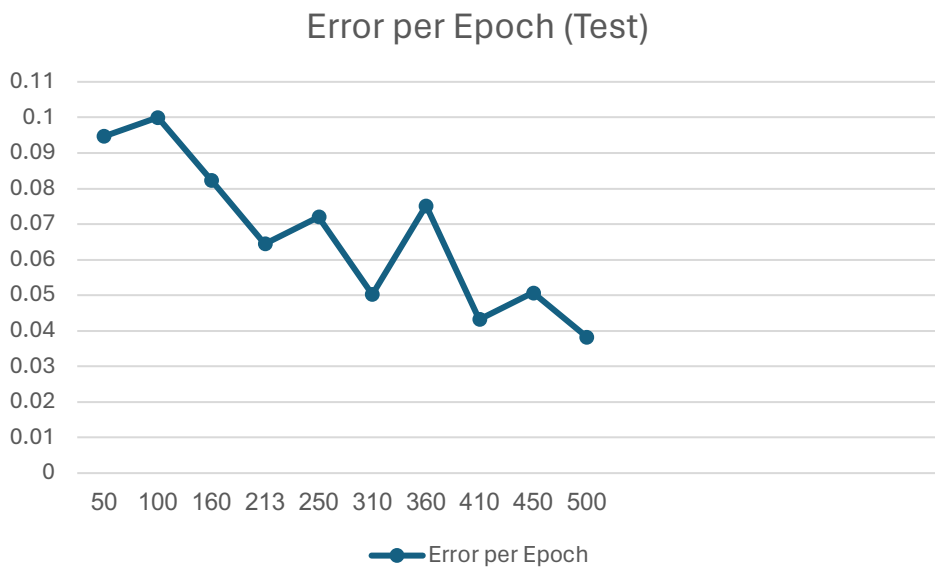
כדי לקבל תמונה ויזואלית של תהליך הלמידה, נייצר גרף המתאר את ערכי פונקציית השגיאה עבור קבוצת האימון וקבוצת המבחן לאורך ה-Epochs. גרף זה יאפשר לנו לעקוב אחר קצב ההתכנסות של המודל, ולזהות תופעות כמו Overfitting (התאמת יתר לנתוני האימון) או Underfitting (התאמה לא מספקת).

להלן גרף השגיאה עבור נתוני האימון (גרף 1):



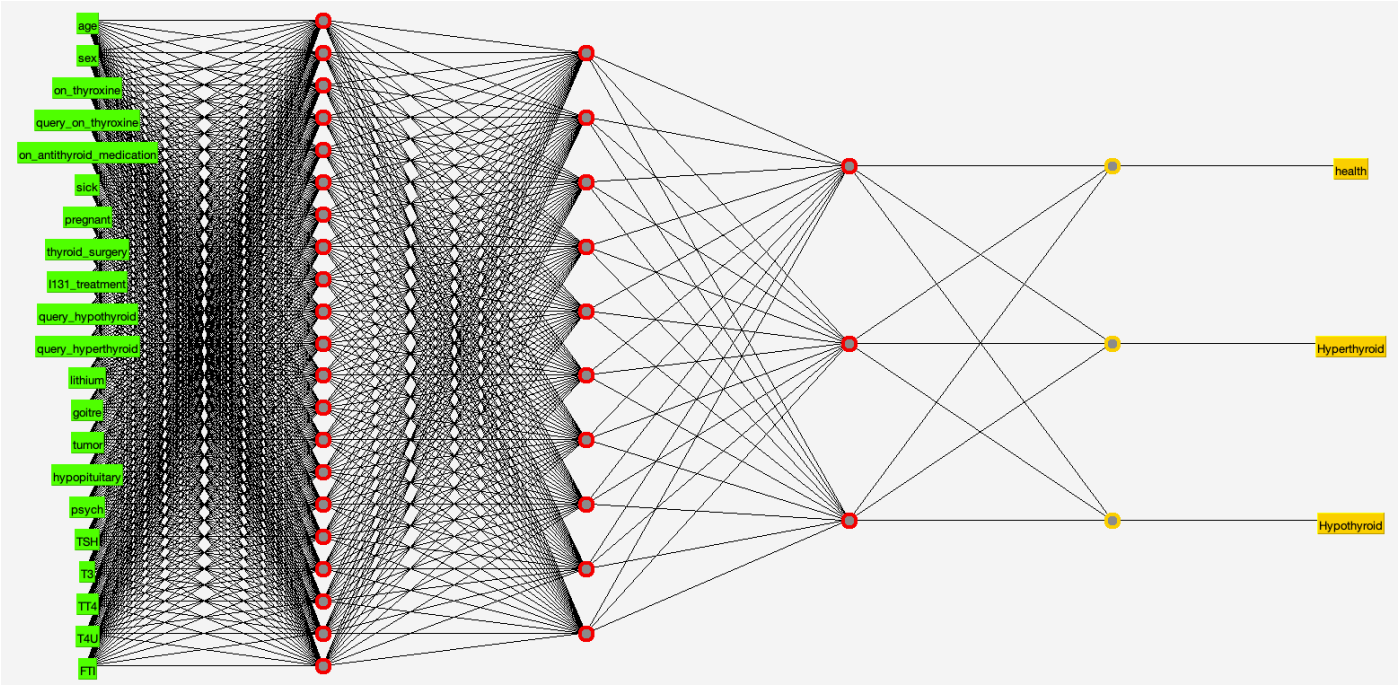
Epoch	Error per Epoch
23	0.0473526
49	0.04733
80	0.0473148
111	0.036707
139	0.0315452
176	0.0302462
206	0.029536
234	0.029036
265	0.0281957
302	0.0266324
350	0.0242792
395	0.0225784
428	0.0205658
462	0.0231116
480	0.0200558
500	0.0195229

להלן גרף השגיאה עבור נתוני המבחן:



Epoch	Error per Epoch
50	0.0947
100	0.1
160	0.0823
213	0.0645
250	0.072
310	0.0503
360	0.0752
410	0.0432
450	0.0507
500	0.0383

להלן ויזואליזציה של מבנה רשת הנוירונים:



הערכת ביצועי הרשת

נבחן את הנתונים שהתקבלו עבור גרף השגיאה עבור נתוני האימון (ערכי פונקציית ה-Loss (MSE) עבור קבוצת האימון לאורך ה-Epochs):

- ה-MSE יורד בצורה משמעותית מערך של כ-0.0474 ב-Epoch 0 לכ-0.0195 ב-Epoch 500.
- ערכי ה-MSE יורדים בצורה עקבית עם ההתקדמות באימונים. דבר זה מצביעה על כך שהרשת לומדת ומשפרת את יכולותיה.
- הירידה החדה ביותר מתרחשת במאה ה-Epochs הראשונים, ולאחר מכן קצב השיפור הופך לשטוח יותר.
- העובדה שה-MSE ממשיך לרדת עד ה-Epoch האחרון מרמזת על כך שהמודל טרם הגיע לנקודת הרוויה (Plateau), ויתכן שהמשך האימונים היה מוביל לשיפור נוסף.

נבחן את הנתונים שהתקבלו עבור גרף השגיאה עבור נתוני המבחן (ממוצע ערכי השגיאות (MAE) עבור קבוצת המבחן לאורך ה-Epochs):

- ציר ה-x מייצג את מספר ה-Epochs (נקודות מרכזיות: 50, 100, 160, ..., 500).
- ציר ה-y מייצג את ערכי ה-MSE של קבוצת המבחן.
- הגרף מראה ירידה משמעותית ב-MSE, מ-0.0947 ב-Epoch 50 ל-0.0383 ב-Epoch 500, עם מספר תנודות לאורך הדרך.
- הפער בין ה-MSE על קבוצת האימון וקבוצת המבחן מצטמצם (פער גדל עשוי להעיד על בעיית Overfitting)

משני הגרפים עולה כי המודל משתפר בהדרגה עם התקדמות תהליך הלמידה, כאשר ה-MSE יורד הן עבור קבוצת האימון והן עבור קבוצת המבחן. הפער בין ה-MSE של קבוצת האימון לזה של קבוצת המבחן אינו גדול מדי, מה שמרמז על כך שהמודל מצליח להכליל היטב גם לנתונים חדשים, ואינו סובל מבעיית Overfitting חמורה. עם זאת, הערך הסופי של MSE על קבוצת המבחן (0.0383) מעט גבוה יותר מאשר על קבוצת האימון (0.0195229), מה שעשוי להעיד על פער מסוים בין יכולת המודל "להתאים" לנתוני האימון לעומת הביצועים על נתונים חדשים. בסיום האימון, המודל משיג דיוק של 95.45% על קבוצת המבחן (3272 דגימות מתוך 3428 מסווגות נכונה), מה שמצביע על רמת ביצועים גבוהה למדי במשימת הסיווג.

להלן סיכום התוצאות:

```
=== Summary ===

Correctly Classified Instances        3272               95.4492 %
Incorrectly Classified Instances       156                4.5508 %
Kappa statistic                          0.5931
Mean absolute error                      0.034
Root mean squared error                  0.1624
Relative absolute error                 36.3811 %
Root relative squared error             75.8703 %
Total Number of Instances             3428

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.993    0.488    0.963      0.993    0.978      0.642    0.955     0.992     health
                 0.345    0.006    0.744      0.345    0.471      0.490    0.937     0.587     Hyperthyroid
                 0.753    0.004    0.809      0.753    0.780      0.776    0.993     0.725     Hypothyroid
Weighted Avg.    0.954    0.453    0.948      0.954    0.947      0.637    0.955     0.966

=== Confusion Matrix ===

    a    b    c   <-- classified as
 3156   11   11 |    a = health
  114   61    2 |    b = Hyperthyroid
    8   10   55 |    c = Hypothyroid
```

2.4 – דיווח על מקרי סיווג שגויים

נתמקד בזיהוי ובחינה מעמיקה של דגימות ספציפיות מתוך קבוצת המבחן שעבורן המודל שלנו ביצע סיווג שגוי. ניתוח המקרים החריגים הללו חיוני להבנת מגבלות הרשת, זיהוי דפוסים בטעויות ובניית אסטרטגיות לשיפור עתידי.

להלן מטריצת הבלבול (Confusion Matrix) שהתקבלה מהערכת המודל על קבוצת המבחן:

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 3156   11   11 |    a = health
  114   61    2 |    b = Hyperthyroid
    8   10   55 |    c = Hypothyroid
```

ניתוח שגיאות הסיווג

הבעיה הבולטת ביותר היא הסיווג השגוי של 114 מקרי "Hyperthyroid" כ-"health". זהו כשל משמעותי שכן הוא עלול להוביל לאי-מתן טיפול נחוץ למטופלים הסובלים מתפקוד יתר. גם במקרים של "Hypothyroid" ו-"health" ישנן טעויות לא מבוטלות, אם כי בשכיחות נמוכה יותר:

- 8 מקרי "Hypothyroid" סווגו בטעות כ-"health" ו-10 כ-"Hyperthyroid".
- 11 מקרי "health" סווגו בטעות כ-"Hyperthyroid" ו-11 כ-"Hypothyroid".

טעויות אלו מעידות על קושי של המודל לתפוס את ההבדלים העדינים בין מצבי התפקוד השונים, ולהפריד באופן חד משמעי בין המחלקות.

גורמים אפשריים לשגיאות הסיווג

הגורם העיקרי לשגיאות הוא ככל הנראה ייצוג לא מאוזן של המחלקות במדגם - 92.5% מהדגימות שייכות למחלקת "health", בעוד שרק 5.2% ל-"Hyperthyroid" ו-2.3% ל-"Hypothyroid". הדבר מקשה על המודל ללמוד כראוי את המאפיינים של שתי המחלקות המיעוט.
בנוסף, ייתכן שיש חפיפה במאפיינים של חלק מהמקרים בין המחלקות השונות, מה שמייצר אזורים של אי-ודאות שבהם המודל מתבלבל.
גם מגבלות בייצוג או בבחירה של המאפיינים עלולות לתרום לטעויות הסיווג - אם המידע הנתון אינו מספק או אינו אינפורמטיבי מספיק, המודל עשוי להתקשות לבצע את ההפרדה הנדרשת בין המחלקות.

2.5 – ניתוח התוצאות והסקת מסקנות

מגבלות ונקודות עיוורון של המודל

- ייצוג לא מאוזן של המחלקות: העדפת סיווג דגימות כ-"Healthy" עקב חוסר איזון במדגם.
- חפיפה במאפיינים בין המחלקות: קושי להבחין בין מצבי תפקוד שונים של בלוטת התריס.
- מגבלות בייצוג המאפיינים: מידע נתון לא מספק לצורך סיווג מדויק.

המלצות לשיפור

על אף הדיוק הגבוה של 95.45% על קבוצת המבחן, ניתוח מקרי הסיווג השגויים חושף בעיה משמעותית בזיהוי מצבי תפקוד יתר של בלוטת התריס. כדי לשפר את הביצועים, ניתן לנסות את הגישות הבאות:

- איסוף של יותר דגימות מהמחלקות המיעוט לקבלת ייצוג מאוזן יותר.
- שימוש בטכניקות לאיזון המחלקות, כגון Over/Under-Sampling.
- הנדסת מאפיינים או מציאת מאפיינים אינפורמטיביים נוספים.
- ניסוי של ארכיטקטורות חלופיות לרשת או אלגוריתמים אחרים.

לקחים ומסקנות

תהליך פיתוח מודלים מבוססי למידת מכונה כולל למידה מהטעויות ושיפור מתמיד. ניתוח שגיאות והבנת הסיבות להן הם קריטיים לשיפור המודל ולביצועיו. באמצעות תהליך איטרטיבי של ניתוח טעויות ושיפור מתמיד, ניתן לשפר את ביצועי המודל ולהבטיח סיווג מדויק ואמין יותר. להלן לקחים ומסקנות לגבי שיפור המודל.

איסוף מידע נוסף ושינוי ייצוג המאפיינים:

- איסוף מידע נוסף: ייתכן שיש צורך להוסיף מאפיינים נוספים הנתונים לתוך המודל, כגון היסטוריה רפואית מפורטת יותר, נתונים גנטיים או מידע נוסף על סימפטומים נוספים אשר לא נכללו במודל הנוכחי.
- שיפור ייצוג המאפיינים: ייתכן שיש צורך בבחינה ובשיפור של המאפיינים הקיימים, או בהוספת מאפיינים חדשים שיכולים לשפר את יכולת הסיווג של המודל. למשל, שימוש בטכניקות של Feature Engineering כדי ליצור מאפיינים חדשים ומשמעותיים יותר..

שינויים בארכיטקטורה של המודל ובתהליך האימון:

- שינוי בארכיטקטורה: הוספת שכבות נסתרות נוספות או הגדלת מספר הנוירונים בכל שכבה עשויים לשפר את יכולת המודל ללכוד קשרים מורכבים יותר בין המאפיינים. (יש לציין שכבר נבדקו מספר רב של וריאציות ושינויים בארכיטקטורת הרשת שלא הוזכרו כאן).
- שימוש בפונקציות הפעלה מתקדמות או מתאימות יותר לסט הנתונים: מעבר לפונקציות הפעלה כמו ReLU לשיפור קצב ההתכנסות.
- איזון המדגם: שימוש בטכניקות של איזון מדגם, כגון oversampling של מחלקות המיעוט או undersampling של המחלקה הדומיננטית, יכול לשפר את יכולת המודל לסווג נכון את מחלקות המיעוט.
- שיפור תהליך האימון: התאמת ה-Hyperparameters, כמו קצב הלמידה וגודל ה-Batch, עשויה לשפר את הביצועים של המודל.

במהלך הפרויקט הוכחנו כי רשת ה-Multilayer Perceptron מסוגלת לסווג בעיות בתפקוד בלוטת התריס ברמה גבוהה של דיוק, אך ישנם תחומים לשיפור, במיוחד בסיווג מחלקות המיעוט. השיפורים הנדרשים יוכלו להוביל למודל חזק ומדויק יותר, המצליח לסווג נכון את כל סוגי המקרים הרפואיים הקשורים לבלוטת התריס.

סיכום

לסיכום, על אף הביצועים המבטיחים של המודל, יש צורך בשיפורים ממוקדים כדי להתמודד עם האתגר של זיהוי מדויק של מקרי תפקוד יתר. עם זאת, התוצאות מדגימות את הפוטנציאל של רשתות נוירונים במשימות מורכבות של סיווג מחלות, ומניחות בסיס מוצק להמשך המחקר והפיתוח בתחום.

ניתוח תוצאות השוואתי והסקת מסקנות של תוצאות ממ״ן 21 וממ״ן 22.

סיכום פרויקט – ממ"ן 21: חיזוי סיווג תפקוד בלוטת התריס באמצעות טכניקות כריית מידע

בפרויקט זה חזו את סיווג תפקוד בלוטת התריס באמצעות אלגוריתמי כריית מידע, תוך שימוש בנתונים רפואיים וטיפול בערכים בעייתיים. הנתונים עובדו ונוקו, ולאחר מכן יושמו אלגוריתמי CART ו-C4.5 לסיווג המצבים. תוצאות הניתוח הצביעו על דיוק של 89.86% עבור CART ו-90.27% עבור C4.5, כאשר C4.5 הציג ביצועים מעט טובים יותר. מסקנות הפרויקט מדגישות את הצורך באיזון המדגם ושיפור ייצוג המאפיינים להמשך שיפור הביצועים.

תוצאות:

- דיוק CART: 89.86%
- דיוק C4.5: 90.27%
- סטטיסטיקת Kappa:
- CART: 0.5455
- C4.5: 0.5801
- מורכבות העץ:
- CART: 37 צמתים עלים, גודל כולל של 73
- C4.5: 67 צמתים עלים, גודל כולל של 106

סיכום פרויקט – ממ"ן 22:

חיזוי סיווג תפקוד בלוטת התריס באמצעות ניתוח אשכולות (Cluster Analysis)

בפרויקט זה נעשה שימוש בניתוח אשכולות (cluster analysis) לזיהוי קבוצות של מטופלים עם תסמינים ומאפיינים דומים. האלגוריתם K-Means נבחר לביצוע ניתוח האשכולות בהתבסס על התאמתו למטרות הפרויקט ולסט הנתונים הקיים.

תוצאות:

- מספר האשכולות (K): 3
- דיוק הסיווג: 74.07%

האלגוריתם הצליח לחלק את הנתונים לשלושה אשכולות המייצגים תפקוד תקין, תת פעילות ויתר פעילות של בלוטת התריס, אך שיעור השגיאה של 25.93% מצביע על כך שיש מקום גדול לשיפור. ניתוח ההתפלגות הראה כי רוב הנבדקים הבריאים סווגו נכונה, אך ישנה חפיפה מסוימת בין תכונות של נבדקים עם תת פעילות ויתר פעילות בלוטת התריס.

חיזוי סיווג תפקוד בלוטת התריס באמצעות רשת נוירונים מלאכותית

בפרויקט זה נבחנה רשת נוירונים מלאכותית מסוג MLP (Multilayer Perceptron) לסיווג בעיות בתפקוד בלוטת התריס, תוך התמקדות בארכיטקטורה, אופטימיזציה והערכת ביצועים. ארכיטקטורת הרשת כללה שכבת קלט עם 21 נוירונים, שתי שכבות נסתרות עם 10 נוירונים כל אחת, ושכבת פלט עם 3 נוירונים לסיווג hypothyroid, hyperthyroid ו-healthy. המודל השיג דיוק של 95.45% על קבוצת המבחן, אך זוהו שגיאות סיווג משמעותיות במחלקות המיעוט. המגבלות כללו ייצוג לא מאוזן של המחלקות וחפיפה במאפיינים, מה שמדגיש את הצורך בשיפור ייצוג המאפיינים ואיזון המדגם. תהליך איטרטיבי של למידה מהטעויות ושיפור מתמיד יוכל להוביל למודל מדויק ואמין יותר.

תוצאות:

- דיוק על קבוצת המבחן: 95.45%
- סטטיסטיקת Kappa: 0.5931
- מקרי סיווג שגויים: 114 מקרים של "Hyperthyroid" שסווגו כ-"Healthy".

ניתוח השוואתי

דיוק הסיווג

- CART ו-C4.5 עם 89.86% ו-90.27% בהתאמה.
- MLP: 95.45%
- K-Means: 74.07%

רשת ה-MLP הציגה דיוק גבוה יותר מאלגוריתמי עץ ההחלטות וניתוח האשכולות.

סטטיסטיקת Kappa

- CART: 0.5455
- C4.5: 0.5801
- MLP: 0.593

MLP הציגה סטטיסטיקת Kappa גבוהה יותר, מה שמצביע על הסכמה טובה יותר בין הסיווגים בפועל והחזויים.

מדדי ביצועים נוספים

MLP הציגה שגיאות ממוצעות נמוכות יותר (Mean absolute error, Root mean squared error) ומדדי ROC ו-PRC גבוהים, מה שמצביע על ביצועים טובים יותר בסיווג.

מורכבות המודל:

- CART ו-C4.5: עצים פחות מורכבים.
- K-Means: דורש קביעה מראש של מספר האשכולות.
- MLP: מורכב מאוד, עם שתי שכבות נסתרות.

MLP היה המודל המורכב ביותר, אך השיג ביצועים טובים יותר.

סיכום

השוואת תוצאות הפרויקטים מראה שרשת הנוירונים המלאכותית (MLP) השיגה דיוק גבוה יותר מאלגוריתמי עץ ההחלטות וניתוח האשכולות. ה-MLP הציגה גם סטטיסטיקת Kappa גבוהה יותר ושגיאות ממוצעות נמוכות יותר, אך הייתה מורכבת יותר ודרשה משאבי חישוב רבים יותר. למרות יתרונות הביצועים של ה-MLP, כל המודלים התקשו בסיווג מחלקות המיעוט, מה שמדגיש את הצורך בשיפור ייצוג המאפיינים ובאיזון המדגם. ניתוח האשכולות באמצעות K-Means הציג דיוק נמוך יחסית ושיעור שגיאה גבוה, מה שמעיד על כך ששיטה זו פחות מתאימה לחיזוי במצב הנתון. עם זאת, ניתוח האשכולות סיפק תובנות חשובות לגבי מבנה הקבוצות והמאפיינים ההורמונליים המרכזיים. בסך הכל, רשת הנוירונים הראתה יתרונות משמעותיים בביצועים, אך יש לשקול את מורכבותה ודרישות החישוב בבחירת האלגוריתם המתאים.