תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**סמינריון אפליקציות של כריית נתונים**

**Baby Go**

מגישים : נאור גבאי - 312132830

שלומי אפרגן - 315727339

**תוכן עניינים**

פרק 1 – מבוא

רקע ופתיחה........................................................................................................................................................................3

ראייה עסקית......................................................................................................................................................................3

כריית נתונים.......................................................................................................................................................................4

פרק 2 – הבעיה העסקית

תיאור הבעיה במונחים עסקיים..............................................................................................................................................5

למי צריך את הפתרון של בעיה מלאכותית...............................................................................................................................5

פרק 3 – בינה מלאכותית

החיבור לכריית מידע ולבינה המלאכותית................................................................................................................................6

פרק 4 – הנתונים

הנתונים...............................................................................................................................................................................7

קורלציה..............................................................................................................................................................................7

רגרסיה לינארית...................................................................................................................................................................8

גרף עמודות..........................................................................................................................................................................9

דיאגרמת עוגה.....................................................................................................................................................................10

פרק 5 – המודלים

האלגוריתמים בהם השתמשנו..............................................................................................................................................11

האלגוריתם הנבחר..............................................................................................................................................................12

המחשה ויזואלית................................................................................................................................................................13

פרק 6 – מימוש האלגוריתם

שלבי העבודה.....................................................................................................................................................................14

פרק 7 – הערכת המודל

הערכת מסווגים.................................................................................................................................................................15

מדדי איכות – הפרמטרים...................................................................................................................................................15

מדדי איכות -תוצאות עבור כל מסווג...................................................................................................................................15

מטריצת הבלבול – תצוגה ויזואלית....................................................................................................................................16

ביבליוגרפיה

ביבליוגרפיה....................................................................................................................................................................19

**פרק 1 - מבוא**

**רקע ופתיחה**

אין בחיים דבר חשוב יותר להורים מאשר הילדים בעיקר כאשר הם פעוטים , ולכן הורים רבים מוטרדים משאלות רבות כגון :

האם התפתחותו המוטורית של הפעוט תקינה?

האם התפתחותו הקוגניטיבית תקינה?

האם התפתחותו הרגשית תקינה?

האם מצבו הבריאותי תקין ומתפתח כראוי ?

האם התזונה שלו נכונה ?

וכמובן עוד שאלות נוספות אשר לפעמים נותרות ללא מענה.

קשה מאוד להשיב לשאלות הללו, כיוון שאין להן תשובות מדויקות , התפתחותו של התינוק תלויה בגורמים שונים, אשר משתנים בחלקם לאורך תקופת הגדילה בצורה אינדיבידואלית בין כל תינוק ותינוק , ואף קשורים בקשר ישיר לגורם התורשתי.

אך עם זאת יש חשיבות רבה גם לגורמים נוספים , אשר יכולים לסייע במתן חיזוי וידע מקדים להתפתחותו התקינה של הפעוט.

המעקב אחר התפתחות ילדינו הצומחים הוא דבר אותו אנחנו לומדים מהרגע הראשון בו אנחנו יוצאים מבית החולים, וכבר בביקור הראשון בטיפת חלב מלמדת האחות שיש לעקוב אחרי המשקל והגובה של הילדים, כדי לבחון האם ההתפתחות שלהם מתנהלת כראוי.

בדרך כלל נהוג למדוד את המשקל ואת הגובה של הילדים מדי חודש בשנתם הראשונה, ובכל כמה חודשים לאחר מכן, אבל זה דבר שתמיד כדאי לעקוב אחריו.

אפליקציית Baby Go מאפשרת להורים לעקוב אחר קצב גדילת הילדים (בגילאי 3-0), לבדוק ולתעד לאורך זמן היכן ממוקמים הילדים על עקומת הגדילה ביחס לאוכלוסייה , בנוסף האפליקציה מתבססת על מס' פרמטרים נוספים מלבד גובה ומשקל , אשר משקפים את מצבו ההתפתחותי של הפעוט ובמקרה הצורך אף תתריע בפני ההורים על בעיה התפתחותית אשר קיימת כיום או תיגרם בעתיד.

**ראיה עסקית**

כיום , ניתן ללמוד ולחקור באופן עצמאי על התפתחותו של הפעוט על ידי קריאת מאמרים , התייעצות עם רופאים , קריאת ספרי מדע ורפואה, התייעצות עם בני משפחה ,חברים ועוד.

אך עם זאת לא קיים מענה או פתרון ממשי אשר עונה על השאלות הרבות באופן חד משמעי ,ולכן הורים רבים נותרים עם סימני שאלה וחששות רבות לגבי גדילה והתפתחות תקינה של ילדיהם.

מבחינה עסקית , שימוש בפתרון טכנולוגי עבור הבעיה יהווה כמקור פתרון ראשון ,מהיר ממוקד ויעיל , אשר ישמש הורים רבים , בעיקר את הטריים אשר הינם אובדי עצות בנושא.

תינתן אפשרות בפני ההורים לתכנן ללמוד ואף לעקוב אחר קצב הגדילה וההתפתחות של פעוטם ובמקרה הצורך להתמקד בבעיה אשר האפליקציה חוזה בעתיד.

מטרתה העיקרית של אפליקציית Baby Goהיא הענקת ביטחון וידע נרחב עבור הורים אשר מעוניינים ללמוד ולעקוב אחר קצב ההתפתחות של בנם ולדאוג שהיא אכן תקינה ומתפתחת כראוי , האפליקציה בודקת מספר פרמטרים : מגדר , משקל , גובה , היקף ראש , גיל , כמות ברזל , שעות שינה , ויטמינים , אשר מהווים בסיס מדעי לקצב הגדילה וההתפתחות של הפעוט.

בעזרת האפליקציה ניתן יהיה לצפות ולחזות מצבי התפתחות אשר יחולקו ל3 קבוצות עיקריות :

1. תינוק בריא
2. תינוק בעל בעיה קוגניטיבית
3. תינוק בעל בעיה פיזיולוגית

**כריית נתונים**

הרעיון העיקרי של המערכת שלנו היא לגרום למכונה ללמוד נתונים רבים ((Data , ועל סמך הנתונים לתת מענה ישיר עבור פתרון הבעיה הדרושה בהתאם לנתונים שלמדה המערכת.

המערכת מתבססת על איסוף מידע נרחב מהרשת הכולל מס' רחב של פרמטרים אשר נבדקו בקפידה רבה , בעזרת הפרמטרים יוכל המשתמש לעקוב ולחזות בקצב הגדילה של התינוק ואף לוודא שהוא אכן מפתחת כראוי , ובמידת הצורך תתריע המערכת בפני המשתמש על בעיית התפתחות שעלולה להיווצר בעתיד.

מתוך איסוף המידע הנרחב המערכת תשמור מדדים תקינים ותסווגם כבריא (על פי בדיקה מעמיקה של הנתונים) , בנוסף המערכת תשמור מדדים שאינם תקינים ותסווגם כלא בריא בהתאם לשני קבוצות : בעיה קוגניטיבית , בעיה פיזיולוגית (על פי בדיקה מעמיקה של הנתונים בהתאם לסוג הבעיה).

המשתמש נכנס למערכת , ומזין בזמן אמת אחר הפרמטרים הדרושים למערכת וממתין לקבלת תשובה.

המערכת מבצעת מודל חיזוי על סמך נתוני המשתמש (קבוצת הבקרה) , ונתוני המערכת הקיימים (קבוצת האימון) ולבסוף מסווגת האם הפרמטרים של התינוק אכן תקינים והתינוק אכן מתפתח כראוי , במידה ולא , המערכת תסווג את הפעוט לקבוצת הבעיות הרלוונטיות עבורו.

המערכת תשמור את הנתונים של המשתמש , ויהיה באפשרותו לעקוב באופן סדיר ורציף אחר הדיווחים וקצב ההתפתחות מעת לעת.

סדר פעולות – עבור המשתמש במערכת:

* המשתמש נכנס לאפליקציה ומזין את הנתונים הדרושים למערכת.
* המערכת מריצה אלגוריתם אשר מבצע סריקה מקיפה על הנתונים של המשתמש בהתאמה לבסיס הנתונים הקיים במערכת.
* המערכת מסווגות את הנתונים בעזרת האלגוריתם ולבסוף מציגה בפני המשתמש פלט סופי עבור מצבו של הפעוט , תינתן אפשרות סיוע לשיפור המצב הקיים במידה ומצבו ההתפתחותי של הפעוט אינו תקין.

**פרק 2 – הבעיה העסקית**

**תיאור הבעיה - במונחים עסקיים**

הורים טריים רבים מדווחים כי ברגע שהם עוזבים את בית החולים עם התינוק החדש, הם משוכנעים שכל פעולה שלהם היא שגויה או עלולה לסכן את בריאות העולל, אפילו כשמדובר במטלות בסיסיות כמו החלפת חיתול.

מעבר לעובדה שאימהות רבות חוות שינויים הורמונליים שגורמים להן לחוש חרדה רבה יותר, מדובר בתפקיד חדש לחלוטין שההורים הטריים מתנסים בו לראשונה, ולכן הפתרונות הקיימים עבור ההורים הם:

ראשית , להתייעץ עם רופא משפחה שאף לא תמיד יכול לקבל אותם בזמן הנוכחי ולכן ידרשו להמתין או לקבוע תור בהתאם.

שנית , להתייעץ עם בני משפחה שלא תמיד הם בעלי ידע עבור פתרון הבעיה , ואף יכולים להטעות את ההורים בנוגע לשאלותיהם.

שלישית , קריאת מאמרים או כתבות שאף לא תמיד מבוססים על סמך פתרון מדעי אלא על טענות או קונספירציות שאינן מבוססות ובדרך כלל משמשות כהשערות מצדו של הכותב , ולכן למרות הזמינות וקלות החיפוש , קשה לדעת מה מתוך המידע ברשת אכן אמין ומה לא.

לסיכום , קיימים מגוון פתרונות שונים עבור הורים אשר מעוניינים לצבור ידע לגידול והתפתחות נכונה של בנם , אך אין פתרון חד משמעי אשר מהווה כמוקד ראשי ומבוסס לפתרון הבעיה.

**למה צריך את הפתרון של בינה מלאכותית?**

השימוש בבינה מלאכותית יכול לשפר את עבודת המחקר ואף לייעל אותה בעת חיפוש פתרון לבעיה.

בינה מלאכותית מאפשרת בקרה פנימית ויכולת זיהוי כשלים שעד היום לא ניתן היה לזהות , להוסיף יותר מידע (בכמויות גדולות) , ולבצע זאת בפחות זמן ומשאבים. כל זאת, ללא פגיעה באיכות העבודה (ולפעמים אף בשיפורה).

בינה מלאכותית היא למעשה שילוב של האינטליגנציה האנושית יחד עם תפעול רובוטי ממוחשב אשר מהווה מענה ישיר לפתרון הבעיה , ובכך מייצרים יכולת ביצוע העולה על גבולות היכולת האנושית.

למעשה, אפשר יהיה לשפר משמעותית את רמת החיזוי המבוססת על מערכי נתונים ובכך רמת הדיוק של האלגוריתם תהיה גבוהה אף יותר מפתרון שאינו מהווה שימוש בבינה המלאכותית.

עבור פתרון הבעיה אשר ניצבת לפנינו ,שימוש בבינה מלאכותית הכרחי ואף יעיל ביותר על מנת לקבל את המענה המהיר והמדויק ביותר בהתאם לכמות המידע הנרחבת אשר נדרשת לפתרון הבעיה.

**פרק 3 – בינה מלאכותית**

**החיבור לכריית מידע ולבינה המלאכותית**

כבר היום חברות, עסקים, ארגונים, מדענים וממשלות משקיעים מאמצים ומשאבים רבים בפיתוח כלים אוטומטיים וטכנולוגיות לצורך כריית מידע , עסקים מסחריים עושים שימוש נרחב בכריית נתונים על מנת לזהות פרופילים של לקוחות, לתקשר איתם ולהתאים להם מסרים, שירותים ומוצרים.

חברות ביטוח ובנקים משתמשים בכריית מידע כדי לזהות הונאות וחריגות, ממשלות ממנפות טכנולוגיית כריית נתונים לזיהוי מוקדם של טרור ופשע ולחיזוק הביטחון הלאומי, ועולם המדע משתמש בטכניקות של כריית נתונים על מנת להגיע לתובנות מדעיות חשובות וחדשות שמקדמות את העולם ואת האנושות כולה.

אפליקציית Baby Go מתבססת בעיקרה על שימוש ב[אלגוריתם](https://he.wikipedia.org/wiki/%D7%90%D7%9C%D7%92%D7%95%D7%A8%D7%99%D7%AA%D7%9D) לצורך סקירה , חקירה , וניתוח מידע הטמון ב[בסיס נתונים](https://he.wikipedia.org/wiki/%D7%91%D7%A1%D7%99%D7%A1_%D7%A0%D7%AA%D7%95%D7%A0%D7%99%D7%9D) קיים , במטרה להסיק ולזהות דפוסי [מסקנות](https://he.wikipedia.org/wiki/%D7%94%D7%99%D7%A7%D7%A9) מהצלבתו (תהליך כריית מידע) , כלומר זיהוי פתרון עבור כל דגימה חדשה מתוך תהליך הלמידה של המכונה מהנתונים הקיימים במערכת.

האפליקציה מכילה כמות גדולה של מידע על מנת לבסס את הפתרון הטוב ביותר עבור כל דגימה חדשה המוכנסת למערכת , ובשל כך הניתוח נעשה באמצעים אוטומטיים , על מנת לקבל את המענה המהיר ביותר עבור המשתמש.

הפלט הסופי של המערכת מציג את 3 המצבים האפשריים למצבו ההתפתחותי של הפעוט :

בריא , בעיה קוגניטיבית ,בעיה פיזיולוגיות.

בנוסף יוצגו הודעות עבור המשתמש על מנת לתקן את המצב הקיים במידת הצורך , דבר אשר ימנע מהמשתמש לחפש פתרונות במקומות אחרים.

**פרק 4 – הנתונים**

רוב הנתונים עליהם מתבסס האלגוריתם הם נתונים בעלי ערכים מספריים אשר נבחרו על פי כמה קריטריונים : גיל (בחודשים) , משקל (בקילוגרמים) , גובה (בסנטימטרים) , היקף ראש (בסנטימטרים) , כמות שעות השינה ( בשעות) , ברזל וויטמין D (מיקרוגרם). מלבד נתון אחד שהוא בעל ערך קטגוריאלי , מין (טקסט).

הקריטריונים עליהם מתבסס האלגוריתם מהווים שדות מפתח על מנת לבצע חיזוי עתידי למצבו של התינוק ואף לשייך אותו ל3 קבוצות עיקריות : בריא , בעל בעיה פיזיולוגית , בעל בעיה קוגניטיבית. כל פרמטר מהווה קשר ישיר לאחת מסוגי הקבוצות עליהם מתבסס האלגוריתם.

איסוף הנתונים הנדרשים לשם אבחון מצבו ההתפתחותי של התינוק נלקחו ממקורות אמינים אשר מפוקחים ומאושרים על ידי משרד הבריאות כגון אתרי : כללית , מכבי , מטרנה , שניידר , המרכז הרפואי הדסה , ועוד רבים נוספים אשר מהווים ביסוס מדעי להתפתחותו של התינוק .

בכל אתר נאספו כמות מחקרים ומאמרים על מנת לבסס את המידע ובנוסף נעשה שימוש במחשבוני גדילה לתינוק על מנת לשקלל את הפרמטרים הנדרשים.

המערכת מכילה כמויות אדירות של מידע המגיע ממקורות שונים. את המידע צריכים לשמור, לנהל, לתחזק ולעבד בקלות וביעילות, תוך השקעה של מינימום משאבים ושימוש בכלים אוטומטיים.

הערכים אשר מהווים את בסיס הנתונים ברובם ערכים מספריים פרט לעמודה בעלת ערכים קטגוריאליים כפי שציינו קודם , על מנת שכול הנתונים יהיו בקנה אחד של טווח ערכים נרמלנו את הערכים לטווח של בין 0 ל1 לפי שיטת MinMax\_Scale.

בנוסף , את העמודה בעלת הערכים הקטגוריאליים המרנו ל0 ו1 על מנת להשתמש באלגוריתמי סיווג אשר יסייעו לנו בזיהוי מקדים של מצבו ההתפתחותי של הפעוט.

ערכים חסרים אינם קיימים במערכת , וזאת משום שבסיס הנתונים הקיים במערכת הוא בסיס נתונים הנלקח מתוך אתרים בעלי אסמכתא רפואית , שבהם כל ערך חושב לפי מחשבון בהתאמה , ולכן לערך חסר לא תינתן האפשרות לחישוב ולא תינתן האפשרות לסיווגו.

**קורלציה**

מדד הקורלציה – מתאר עד כמה הקשר בין שני עמודות תלוי אחד בשני.

הערכים של הקורלציה נעים בין 1 ל 1- , ככול שאנו יותר קרובים ל1 הקשר הוא קשר חזק וחיובי.

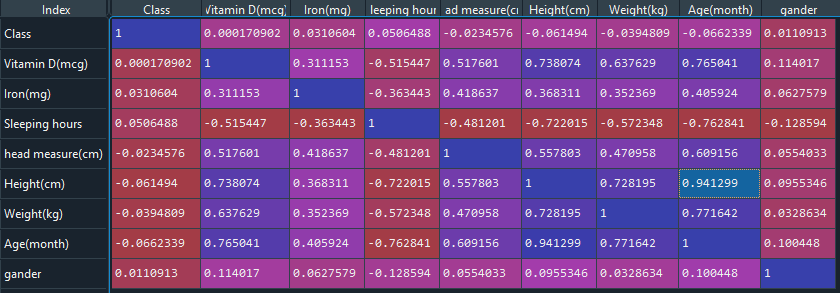
ככול שאנו יותר קרובים ל1- הקשר הוא קשר חזק ושלישי.

ככול שאנו קרובים ל0 הקשר הוא חלש.

כאשר מדדנו את הקורלציה בין כל העמודות (לא כולל עמודת הסיווג) ראינו כי יש קשר חזק וחיובי (קשר ישיר) בין העמודה גיל לעמודה גובה שהוא 0.94 מה שמעיד על כך שיש תלות גדולה מאוד בין שני העמודות ולכן אפשר לפסול אחת מבניהן.

ניתן לראות כי בנוסף קיים קשר חזק וחיובי בין העמודות ויטמין די לבין הגיל אך הקשר החזק ביותר הוא בין העמודות גיל וגובה כפי שציינו קודם.

כאשר נמחק אחת מן העמודות נגרום להפחתת הנתונים , דבר שיגרום לחיסכון ושיפור בזמן הריצה של האלגוריתם , ובנוסף ישפר את ביצועי האלגוריתם (רמת הדיוק).

דוגמא : אלגוריתם Knn הוא אלגוריתם אשר מוגבל במקרה של מספר רב של עמודות ולכן הפחתה של עמודות תגרום לשיפור בביצוע

**רגרסיה לינארית**

הרעיון הוא שכאשר זיהנו תלות לינארית בין שני משתנים , כמו שציינו קודם על ידי חישוב מדד הקורלציה, מצאנו כי בין העמודה גיל לעמודה גובה קיים קשר חזק וחיובי.

על מנת להציג את טיב הקשר הלינארי בין שני המשתנים , בצורה גרפית , נציג את קו הישר הלינארי המייצג בצורה הטובה ביותר את התלות בין שני המשתנים.

נעביר קו לינארי ונחשב את המרחק של כל נקודה מהקו הלינארי שהעברנו , בעזרת נוסחת הרגרסיה הלינארית , ובכך נמצא את הקו הלינארי.

ככול שהקו הלינארי יהיה קרוב יותר לכול הנקודות כך הוא ייצג טוב יותר את התלות מבין שני המשתנים.

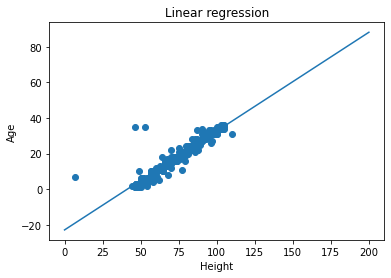
המטרה : לחזות או לצפות בעזרת ערך שמציבים במשוואה כמה נקבל על סמך חישוב המשוואה של הרגרסיה הלינארית.

סוגים של קשר לינארי:

קו לינארי ישיר – ככול שמשקיעים יותר לדוגמא בפרסום כך המכירות יגדלו (קשר חזק וחיובי) כמו בתוצאה שקיבלנו בין העמודה גיל לגובה.

קו לינארי הפוך/שלילי – ככול שמשקיעים יותר לדוגמא בפרסום כך המכירות ירדו.

ניתן לראות כי על פי הגרף , קיבלנו קשר חזק וחיובי בין שני העמודות ( גיל וגובה ) , הקו הלינארי מייצג בצורה טובה את התלות בין שני העמודות , הקו הלינארי מאוד קרוב לנקודות פרט לשני נקודות שהם חריגות ביחס לרוב הנקודות , ובשל כך ניתן להסיק כי אכן קיים קשר חיובי וחזק בדיוק כמו בתוצאה שקיבלנו בחישוב מדד הקורלציה.



**הכרה של בסיס הנתונים במערכת – המחשה וויזואלית (לפני תהליך העיבוד/סיווג דגימה חדשה)**

**גרף עמודות**

ניתן לראות כי גרף העמודות מציג בפנינו את הנתונים המקוריים של בסיס הנתונים הקיים בערכת.

מצבם ההתפתחותי של התינוקות מחולק ל3 קטגוריות כפי שציינו קודם , ולכן ניתק להסיק מן הגרף כי :

תינוקות בריאים – יש סביבות 100 פעוטות אשר מצבם ההתפתחותי מוגדר כתקין ולכן סווגו כבריאים.

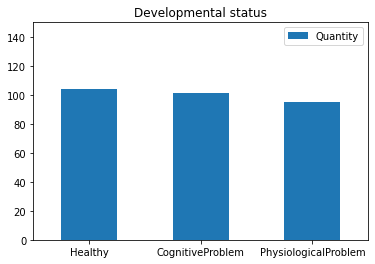
תינוקות בעלי בעיה קוגניטיבית - יש קצת פחות 100 פעוטות אשר מצב ההתפתחותי אינו תקין ועל פי הפרמטרים נקבע כי בעיית ההתפתחות עבורם היא בעיה קוגניטיבית.

תינוקות בעלי בעיה פיזיולוגית – ניתן לראות כי יש פחות פעוטות בעלי בעיה פיזיולוגית מאשר פעוטות בעלי בעיה קוגניטיבית באופן שאינו משמעותי.

חשוב לציין כי אם רוצים את מספרם המדויק של הפעוטות בכל קבוצה וקבוצה נסתכל במשתנה השומר עבור כל קבוצה בנפרד את הכמות המדויקת של הפעוטות בקבוצה (בתוך הקוד).

בעזרת גרף העמודות , אנו מקבלים היכרות כללית עם בסיס הנתונים הקיים במערכת , זאת עוד לפני תהליך העיבוד והסיווג עבור דגימה חדשה שתחושב על ידי הרצת האלגוריתם (קבוצה הבקרה) מתוך בסיס הנתונים הקיים (קבוצת האימון).

היתרון של הגרף הוא נתינת תמונת מצב מקיפה אשר מהווה אומדן כללי לגבי הנתונים הקיימים במערכת , בצורה מובנת וקלה לשימוש.



**דיאגרמת עוגה**

ניתן לראות כי דיאגרמת העוגה מציגה בפינינו את כמות הפעוטות באחוזים , לפי מגדר.

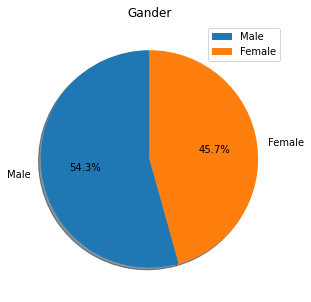
פעוטות אשר מינם מוגדר כזכר – מהווים 54.3% מתוך סך כל הפעוטות אשר נמצאות בבסיס הנתונים.

פעוטות אשר מינם מוגדר כנקבה – מהוות 45.7% מתוך סך כל הפעוטות אשר נמצאות בבסיס הנתונים.

חשוב לציין שדיאגרמת העוגה מהווה יחס כללי באחוזים בין שני קבוצות המגדר , במידה ונרצה את המספר המדויק עבור כל מגדר , נבדוק במשתנה השומר (בתוך הקוד) , את סכומם המדויק של הפעוטות.

בעזרת דיאגרמת העוגה , אנו מקבלים היכרות כללית עם בסיס הנתונים הקיים במערכת , המציג באופן יחסי (באחוזים) את כמות הזכרים והנקבות הקיימים במערכת.

היתרון של הדיאגרמה היא נתינת תמונת מצב מקיפה אשר מהווה אומדן כללי לגבי הנתונים הקיימים במערכת , בצורה מובנת וקלה לשימוש.



**פרק 5 – מודלים**

**האלגוריתמים המסייעים לנו בפתרון הבעיה (סקירה כללית):**

**Knn** - השכנים ה- K הקרובים ביותר, הינו אלגוריתם סיווג.

בשני המקרים משתנה הכניסה תלוי ב- K התצפיות הקרובות ביותר במרחב התכונות של קבוצות במרחב האוקלידי.

יישום האלגוריתם כולל 3 שלבים :

בשלב הראשון, מנרמלים את הנתונים.

בשלב השני, מודדים את המרחק, במרחב של n ממדים, של הנתונים החדשים מהנתונים שעבורם יש לנו תוויות (Labels) תוצאות ידועות מראש.

בשלב השלישי, בוחרים את k פריטי הנתונים הקרובים ביותר - התווית אשר מס' המופעים שלה יופיע ברוב המקרים תייצג את הסיווג עבור הנקודה החדשה.

- **Naïve Base**הינו אלגוריתם סיווג.

עובד על משתנים רציפים ומבוסס על נוסחת בייס , ההנחה ה"נאיבית" שאין תלות בין תכונות האובייקטים המסווגים כאשר כבר ידוע סיווגם. לדוגמא הקשר בין הגובה למשקל הוא קשר תלוי , אך הנחת האלגוריתם מתבססת על כך שהם בלתי תלויים אחד בשני.

האלגוריתם מחשב את ההסתברויות של כל עמודה ומבצע חלוקה לפרמטרים של סיווג , לאחר מכן בודק מהו הרוב לפי ההסתברויות ולוקח את ההסתברות הגבוהה ביותר מבין השניים ובכך יודע לסווג את הפרמטרים שהמשתמש הזין.

**Decision Tree**- הינו אלגוריתם סיווג.

בהינתן קבוצת אימון , נרצה לבנות עץ החלטה מתאים.

נרצה בכל פעם לבחור עמודה , ועבור כל עמודה יש להחליט את הערכים אשר לפיהם אנו מפצלים ולראות מה מקבלים:

1. עלה טהור – כלומר על הנקודות בעלה שייכות לאותו הסוג , ואז העלה יסווג כמו הסוג של אותם נקודות , ולא נמשיך את הפיצול.
2. עלה ריק – העלה יסווג לפי המקרה השכיח בנתונים , ולא נמשיך לפצל את הנתונים.
3. עץ שאינו טהור (מכיל יותר מסוג אחד) – נמשיך לפצל עד שנקבל עליהם טהורים.

מדדי אי טהורות :

1. MSE
2. Gini
3. Entropy

כל אחד ממדדי אי הטהורות מתאר אי טהורות של עלה מסוים , ככול שהמספר של מדד אי הטהורות קטן יותר כך התוצאה טובה יותר.

על מנת לבחור את המסווג בעל התוצאה הטובה ביותר , נבצע הערכת מסווגים על ידי בניית מטריצת הבלבול (Confusion Matrix ( ונקבל עבור כול מסווג באופן ייחודי את מדדי האיכות שלו :Recall , Precision , F-score, Accuracy ועל פי כן נקבע מיהו המסווג בעל התוצאה הטובה ביותר.

האלגוריתם Decision Tree נבחר כאלגוריתם הטוב ביותר מבין השלושה וזאת בשל התוצאות הגבוהות שלו על פי מדדי האיכות ביחס לשאר האלגוריתמים.

בכל הרצה קבוצת הבדיקה משתנה באופן רנדומלי כך שהבדיקה נעשית בכל פעם ופעם על קבוצת בקרה שונה (בעלת ערכים שונים) , אך בכל פעם כאשר מריצים את שלושת המסווגים התוצאה הגבוהה ביותר מתקבלת עבור האלגוריתם עצי החלטה ולכן הוא הנבחר מבין השלושה.

**Decision Tree – האלגוריתם הנבחר (סקירה מעמיקה)**

עץ החלטה הוא מודל חיזוי בתחומי הסטטיסטיקה , המספק מיפוי בין תצפיות לערכים המתאימים עבורן.

עץ החלטה הוא [עץ בינארי](https://www.hamichlol.org.il/%D7%A2%D7%A5_%D7%91%D7%99%D7%A0%D7%90%D7%A8%D7%99) מלא המורכב מצמתי החלטה שבכל אחד מהם נבדק תנאי מסוים על מאפיין מסוים של התצפיות ועלים המכילים את הערך החזוי עבור התצפית המתאימה למסלול שמוביל אליהם בעץ.

למידה מבוססת עץ החלטה היא אחת הטכניקות המוצלחות ביותר ללמידה באמצעות סיווג.

עץ יכול "ללמוד" על ידי פיצול קבוצת המקור לתתי קבוצות, המתבססות על מתן ערך לתכונה. תהליך זה חוזר על עצמו בכל תת-קבוצה באופן רקורסיבי.

**יתרונות**

.1קל להבנה, והסקת מסקנות.

2. לא דורש המון פעולות להכנת בסיס הנתונים.

3.יכול לקבל ערכים מספריים (רציפים , בדידים) , וקטגוריאליים.

4.עלות השימוש בעץ נמוכה (זמן ריצה).

5. פועל היטב עם מערכי נתונים גדולים.

6.ערכים חסרים אינם משפיעים על תהליך בניית העץ ההחלטה במידה ניכרת.

6. בחירת עמודות , האלגוריתם מכיל בתוכו שיטת בחירת עמודות בעלות החשיבות הגבוה ביותר.

האלגוריתם לומד את קבוצת האימון , ובוחן את העמודות הרלוונטיות ביותר על ידי קבוצת האימון.

**חסרונות**

1.קיימת נטייה לעשות Overfitting לבסיס הנתונים.

2. שינויים קטנים בבסיס הנתונים עלולים לגרום לשינויים גדולים בתוצאות העץ.

3. כאשר יש דאטה שאינו מאוזן מבחינת כמויות , חוסר האיזון מתבטא בתוצאת המסווג.

**סיבות לבחירת האלגוריתם Decision Tree**

1. בבסיס הנתונים שלנו קיימים ערכים מספריים רציפים ובדידים ובנוסף גם ערכים קטגוריאליים (במקור) , אלגוריתם עצי החלטה יודע לעבוד אם כל סוגי הערכים ובכך יכולת ההתמודדות שלו מול בסיס הנתונים היא טובה ביחס לשאר האלגוריתמים.
2. בסיס הנתונים שלנו מהווה כ300 דגימות (שורות) , ו9 תכונות (עמודות) , אלגוריתם עצי החלטה מתמודד עם בסיסי נתונים אף גדולים בהרבה מכך ולכן התוצאה היא טובה ביחס לשאר האלגוריתמים.
3. ניתן למדוד את מדדי ביצועי האלגוריתם ותוצאות הסיווג שלו , ולאחר הרצת שלושת האלגוריתמים ניתן לראות באופן מובהק כי תוצאת המסווג של עצי החלטה היא הטובים ביותר עבור בסיס הנתונים שלנו.

**עץ החלטה – המחשה ויזואלית**

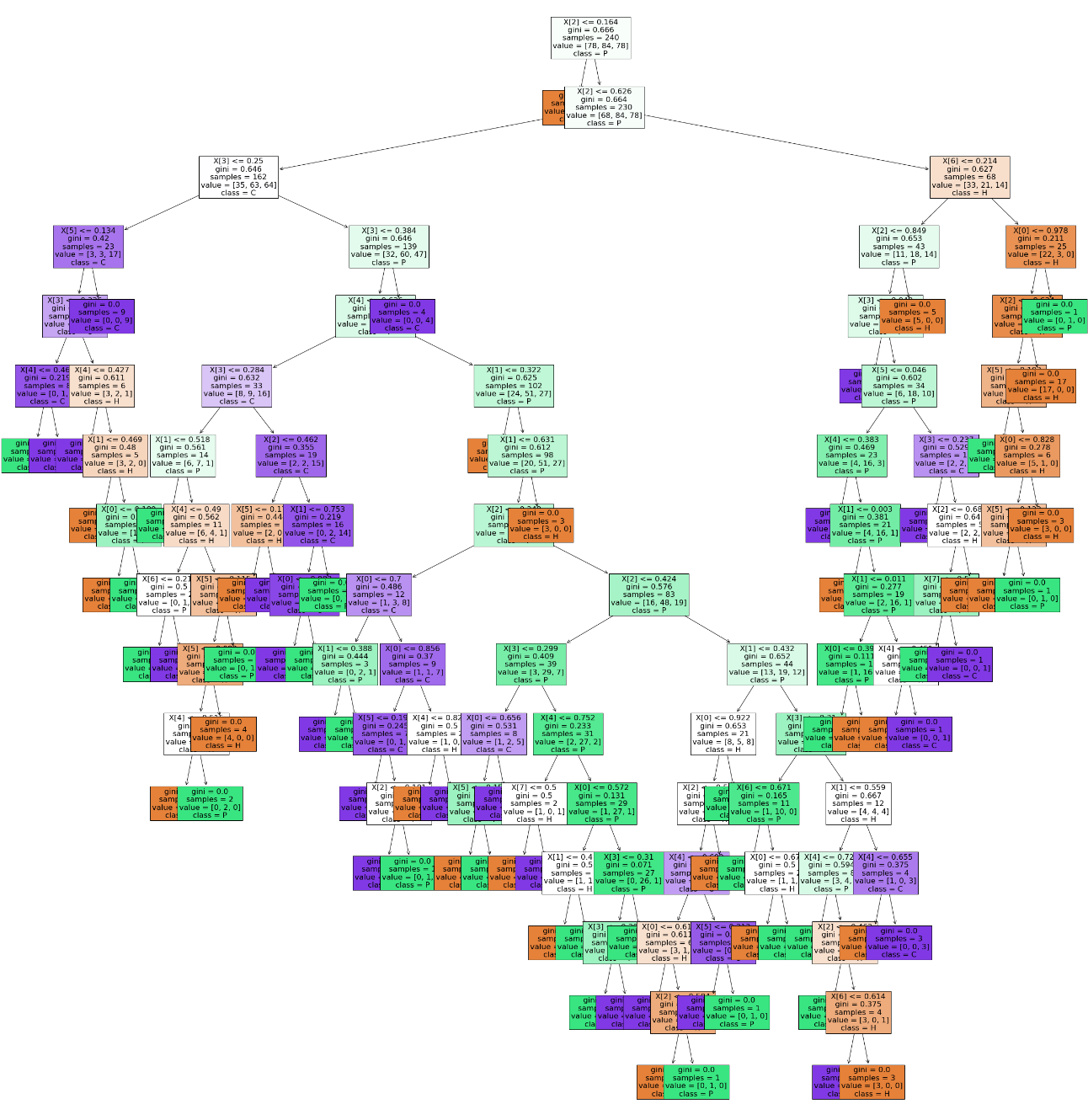
בתמונה ניתן לראות את תהליך הסיווג עבור האלגוריתם עצי החלטה מתחילתו (שורש) ועד סופו.

הפרמטרים ליצירת העץ בעזרת פייתון הם :

1. קביעת עומק העץ , ברירת המחדל היא עומק מלא.
2. שמות התכונות (העמודות) , ברירת המחדל היא שימוש בשמות גנריים כמו שמופיע בתמונה ("("X[0] בראשית הצומת.
3. גודל הגופן אשר יוצג בכל צומת וצומת.
4. צביעת הצמתים כדי לציין את דרגת הרוב לסיווג.

בכל צומת ניתן לראות כי נשמרים מס' ערכים :

1. תוצאת מדד הGini
2. כמות הדגימות
3. כמות הערכים עבור כל סיווג
4. תוצאת הסיווג הסופית
5. העמודה אשר מייצגת את הצומת



**פרק 6 – מימוש האלגוריתם**

**שלבי עבודה – מימוש אלגוריתם**

1. ראשית , אספנו וחקרנו המון מידע מתוך האינטרנט ממגוון אתרים בעלי אסמכתא רפואית אשר מהווים ביסוס מדעי לכל התוכן אשר מוצג באתר.
2. בעזרת מחשבוני גדילה כגון – משקל , גובה , תזונה , חשיפה לאור , שעות שינה , היקף ראש ועוד מגוון פרמטרים , ניתן היה לקבוע איזה סוג בעיית התפתחות (במידה ויש) תיגרם לפעוט לפי הנתונים המחושבים בעזרת מחשבוני הגדילה , ובנוסף קריאה מעמיקה עבור כל סוג בעיית התפתחות ומאפייניה.
3. לאחר איסוף הנתונים , יצרנו בסיס נתונים המכיל כ300 דגימות ו9 עמודות , בהם כל דגימה מהווה פעוט יחיד , וכל עמודה מהווה מאפיין/תכונה של אותו הפעוט.
4. לאחר יצירת בסיס הנתונים , טענו את קובץ הCSV המכיל את כל הדאטה אשר יצרנו באמצעות ספריית pandas.
5. בבסיס הנתונים המקורי שיצרנו , יש עמודה המהווה ערכים קטגוריאליים לכן המרנו אותה לערכים מספריים.
6. על מנת להמחיש ולהכיר את בסיס הנתונים שיצרנו באופן אשר יהיה קל להבנה יצרנו גרף עמודות , ודיאגרמת עוגה , אשר מציגה באופן כללי את היחס בין הנתונים.
7. ביצענו נרמול לבסיס הנתונים.
8. לאחר עיבוד מקדים לבסיס הנתונים , הדאטה שלנו היה מוכן לשימוש.
9. מתוך בסיס הנתונים יצרנו קבוצת אימון וקבוצת בקרה ( החלוקה היא לפי החלוקה האידאלית 80% מהווה קבוצת האימון ו20% מהווה קבוצה הבקרה).
10. הכנה ומימוש של האלגוריתמי סיווג אשר בחרנו : Knn , Naive Bayes , Decision Tree
11. הרצת כל אלגוריתם על בסיס הנתונים שיצרנו בעזרת קבוצה הבדיקה , וחישוב מדדי הדיוק והאיכות עבור כל אלגוריתם סיווג בנפרד.
12. הערכת מסווגים – בדיקה מיהו המסווג בעל התוצאות הטובות ביותר.
13. הצגת מטריצת הבלבול עבור כל מסווג בנפרד.
14. עבור אלגוריתם עצי ההחלטה ( האלגוריתם הנבחר כבעל התוצאות הטובות ביותר ) הצגנו גרף המדמה את העץ במלואו , ואת כל תהליך הסיווג עד לסיומו.
15. בחירת עמודות (feature selection) – על מנת לבחור את העמודות בעלות הקשר החזק ביותר לתוצאת הסיווג חישבנו קורלציה בין כל עמודה לעמודת הסיווג שלה.
16. בנוסף , ביצענו בדיקה על ידי מדד הקורלציה בין כל עמודה לעמודה (לא כולל עמודת הסיווג) וקיבלנו את מטריצת הקורלציה , בעזרת המטריצה ראינו כי העמודה גיל וגובה תלויות אחת בשנייה בקשר חזק וחיובי ולכן ניתן היה למחוק אחת מבין שני העמודות.
17. רגרסיה לינארית – על מנת להמחיש את הקשר החזק והחיובי בין העמודות הצגנו אותם על ידי גרף וקו לינארי המבטא את הקשר הלינארי (התלות) בין שתי העמודות.
18. לבסוף , הפקנו לקחים מתוך כל התהליך והסקנו מסקנות בהתאם לתוצאות שקיבלנו.

**פרק 7 – הערכת המודל**

**הערכת מסווגים**

מתוך בסיס הנתונים , יצרנו קבוצת אימון וקבוצת בקרה.

קבוצת האימון מכילה כ240 דגימות , וקבוצת הבקרה מכילה כ60 דגימות.

לאחר הרצת אלגוריתמי הסיווג , בדקנו את מדדי האיכות של כל אלגוריתם סיווג בנפרד.

**מדדי איכות – פרמטרים**

משמעותו סך כל התוצאות החיוביות שהמסווג קבע כחיוביות precision –

משמעותו סך כל התוצאות החיוביות האמתיות recall –

משמעותו ממוצע משוקלל של רמת הדיוק (המדד עליו אנו מסתמכים כאשר בודקים את רמת הדיוק של האלגוריתם ) – f1 score

משמעותו רמת דיוק כללית (המדד עליו אנחנו לא מסתמכים כאשר בודקים את הרמת הדיוק של האלגוריתם) Accuracy -

**מדדי איכות – תוצאות עבור כל מסווג**

על פי התמונה , ניתן לראות כי תוצאות הסיווג עבור כל נקודה בקבוצת הבדיקה , היא על ידי הרצת האלגוריתם Knn.

לאחר שקיבלנו את תוצאת הסיווג עבור כל נקודה ונקודה , אנו רוצים לבדוק את מדדי האיכות של האלגוריתם ולכן ניתן לראות בתמונה את מדדי האיכות עבור תוצאת האלגוריתם .Knn

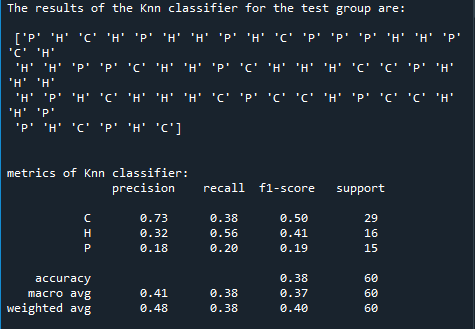
על פי תוצאות מדדי האיכות , ניתן לראות כי המסווג ביצע את הסיווג ברמה סבירה ומטה , ואנו שואפים לתוצאות גבוהות יותר.

יתכן ותוצאת הסיווג המתקבלת נובעת מכמה סיבות אפשריות:

1.ערך הK המוגדר המגביל את התוצאה

2. מס' התכונות מגביל את האלגוריתם (באופן כללי האלגוריתם רגיש למס' רב של עמודות).

3. מבוסס על סמך המרחק האוקלידי ולכן המרחק האוקלידי אינו מדד יעיל למרחק במרחב מממד גבוה.



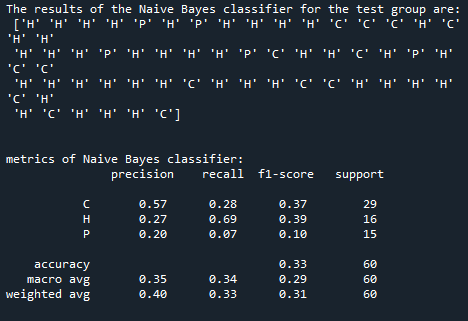
על פי התמונה , ניתן לראות כי תוצאות הסיווג עבור כל נקודה בקבוצת הבדיקה , היא על ידי הרצת האלגוריתם Naïve Bayes.

לאחר שקיבלנו את תוצאת הסיווג עבור כל נקודה ונקודה , אנו רוצים לבדוק את מדדי האיכות של האלגוריתם ולכן ניתן לראות בתמונה את מדדי האיכות עבור תוצאת האלגוריתם Naïve Bayes.

על פי תוצאות מדדי האיכות , ניתן לראות כי המסווג ביצע את הסיווג ברמה פחותה יותר מאשר המסווג Knn , ואנו שואפים לתוצאות גבוהות יותר , לכן מבין השניים נעדיף את המסווג Knn.

יתכן ותוצאת הסיווג המתקבלת נובעת מכמה סיבות אפשריות:

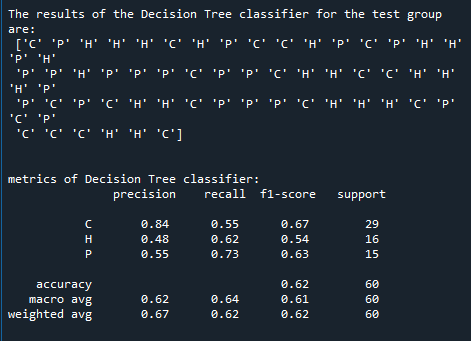
1. ההנחה הנאיבית של האלגוריתם כי אין תלות בין המשתנים.
2. 'בעיית תדר אפס', כאשר הוא מקצה אפס הסתברות למשתנה קטגורי שהקטגוריה שלו במערך נתוני הבדיקה לא הייתה זמינה במערך האימונים.



על פי התמונה , ניתן לראות כי תוצאות הסיווג עבור כל נקודה בקבוצת הבדיקה , היא על ידי הרצת האלגוריתם Decision Tree.

לאחר שקיבלנו את תוצאת הסיווג עבור כל נקודה ונקודה , אנו רוצים לבדוק את מדדי האיכות של האלגוריתם ולכן ניתן לראות בתמונה את מדדי האיכות עבור תוצאת האלגוריתם Decision Tree.

על פי תוצאות מדדי האיכות , ניתן לראות כי המסווג ביצע את הסיווג ברמה טובה יותר מבין שני המסווגים השניים ולכן נעדיף אותו על פני יתר המסווגים.



**מטריצת הבלבול – תצוגה וויזואלית**

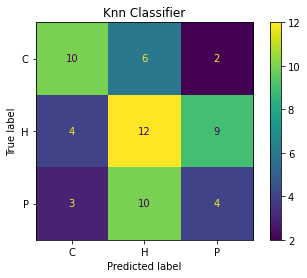
ניתן לראות כי מטריצת הבלבול (confusion matrix) המוצגת היא עבור המסווג Knn.

על פי מטריצת הבלבול , מס' הערכים שהמסווג סיווג ואכן צדק בתוצאת הסיווג הם : 10 + 12 + 4 = 26

מס' הערכים שהמסווג סיווג וטעה בתוצאת הסיווג הם : 6 + 2 +4 + 9 + 3 + 10 = 34

האלכסון הראשי מהווה את תוצאות הסיווג הנכונות בהתאם לקבוצת האימון.

מלבד האלכסון הראשי , כל שאר הערכים מהווים תוצאת סיווג שגויה בהתאם לקבוצת האימון.



ניתן לראות כי מטריצת הבלבול (confusion matrix) המוצגת היא עבור המסווג Naive Bayes.

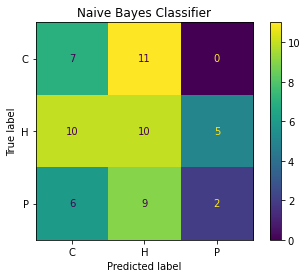
על פי מטריצת הבלבול , מס' הערכים שהמסווג סיווג ואכן צדק בתוצאת הסיווג הם : 2 + 10 + 7 = 19

מס' הערכים שהמסווג סיווג וטעה בתוצאת הסיווג הם : 9 + 6 +5 + 10+ 0 + 11= 41

האלכסון הראשי מהווה את תוצאות הסיווג הנכונות בהתאם לקבוצת האימון.

מלבד האלכסון הראשי , כל שאר הערכים מהווים תוצאת סיווג שגויה בהתאם לקבוצת האימון.

ניתן לראות כי בהשוואה לתוצאות המסווג Knn התוצאות עבור המסווג Naive Bayes היו נמוכות יותר במס' התוצאות שהמסווג סיווג וצדק , ולכן מס' הטעויות של המסווג בתהליך הסיווג היו גדולות יותר בהשוואה למסווג Knn.



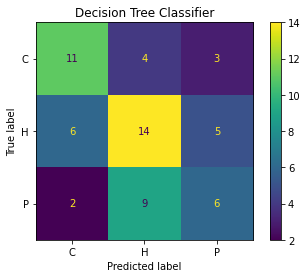
ניתן לראות כי מטריצת הבלבול (confusion matrix) המוצגת היא עבור המסווג Decision Tree.

על פי מטריצת הבלבול , מס' הערכים שהמסווג סיווג ואכן צדק בתוצאת הסיווג הם : 6 + 14 + 11 = 31

מס' הערכים שהמסווג סיווג וטעה בתוצאת הסיווג הם : 9+ 2 +5 + 6 + 3 + 4= 29

האלכסון הראשי מהווה את תוצאות הסיווג הנכונות בהתאם לקבוצת האימון.

מלבד האלכסון הראשי , כל שאר הערכים מהווים תוצאת סיווג שגויה בהתאם לקבוצת האימון.

ניתן לראות כי בהשוואה לשני תוצאות המסווגים Knn ו Naive Bayes תוצאות הסיווג עבור המסווג Decision Tree היו גדולות יותר במס' התוצאות שהמסווג סיווג וצדק , ולכן מס' הטעויות של המסווג בתהליך הסיווג היו נמוכות יותר בהשוואה לשני המסווגים השניים.

**ביבליוגרפיה**

עקומת גדילה בנים – משרד הבריאות <https://www.cdc.gov/growthcharts/data/set1clinical/cj41c019.pdf>

עקומת גדילה בנות - משרד הבריאות <https://www.cdc.gov/growthcharts/data/set1clinical/cj41l020.pdf>

תזונה בריאותית ( ויטמינים ומינרלים) משרד הבריאות <https://www.health.gov.il/PublicationsFiles/nutrition-2020.pdf>

מידע על התפתחויות התינוק לכל הגילאים אתר כללית

<https://www.clalit.co.il/he/your_health/kids/baby/Pages/baby_development.aspx?page=1>

שעות שינה מומלצות לכל הגילאים

<https://www.camoni.co.il/%D7%9B%D7%9E%D7%94-%D7%A9%D7%A2%D7%95%D7%AA-%D7%A6%D7%A8%D7%99%D7%9A-%D7%9C%D7%99%D7%A9%D7%95%D7%9F-%D7%91%D7%9B%D7%9C-%D7%92%D7%99%D7%9C--%D7%9E%D7%93%D7%A8%D7%99%D7%9A>

מחשבון התפתחות התינוק אתר כללית <https://www.clalit.co.il/he/your_health/kids/baby/Pages/infant_growth_calculator.aspx>