חוקי הקשר - Association Rules

נבחר אלגוריתם של חוקי הקשר, נתאר וננתח את האלגוריתם תוך נימוק בחירתנו.

בחרתי באלגוריתם FP-Growth שלמדנו.

אלגוריתם זה בשונה מהאלגוריתם A-priori שלמדנו, עובד בשיטת הפרד ומשול ומוצא קבוצת שכיחויות מבלי לייצר מועמדים, בעזרת שימוש במבנה נתונים בשם FP-Tree (עץ שכיחויות) - ששומר מידע סטטיסטי על הפריטים שבבסיס הנתונים ומהווה שיפור לאלגוריתם A-priori.

האלגוריתם מייצר את העץ ע"י שתי סריקות בלבד של בסיס הנתונים.

האלגוריתם מתבסס על ההנחה שה- Transactions ב-DB ממוינות. (כך גם חלק מהמימושים של A-priori)

את החיסכון הזה הוא מבצע תוך שימוש בעזרת עץ שכיחויות(FP-Tree). כל צומת בעץ מייצג פריט, רמת התדירות של פריט היא <u>ביחס הפוך</u> לעומק שלו בעץ והענף שיוצא ממנו כולל את כל ה-Transactions שקשורות אליו.

פריט בתדירות גבוהה יותר יופיע ברמה נמוכה יותר בעץ, כלומר קרוב יותר לשורש העץ ופריט בתדירות נמוכה יותר יופיע ברמה גבוהה יותר בעץ, כלומר קרוב יותר לעלים.

אלגוריתם זה יעיל כאשר יש קבוצות שכיחות עם מס' פריטים גדול יחסית, מה שמשליך על עץ קומפקטי.

לכל צומת אנו מתחזקים מונה לספירת השכיחויות (מופעים) באותו המסלול. בכדי לקבל את המונה הכולל של פריט בכלל העץ, מתחזקים גם טבלת הצבעות מהפריטים לצמתים בעץ, ומכל צומת יוצרים רשימה מקושרת של הצבעות בין צמתים המייצגים מופעים של הפריט בהסתעפויות שונות. ע"י כך יתאפשר לסרוק את הרשימה ולקבל את המנייה הכוללת של פריט בקבוצה. בניית העץ נעשית מלמעלה-למטה, ושחזור הקבוצות השכיחות וכללי ההקשר החזקים מבוצעים בסיום בניית העץ מלמטה-למעלה, ע"י בניית תתי-עצים מותנים לפריטים השכיחים והתמקדות במסלולים המובילים אליהם.

יתרונות

1. יעיל בהשוואה לאלגוריתם A-Priori
2. מסתפק בסריקת ה DB פעמיים בלבד

חסרונות

1. לא יעיל במקרה של מס' קטן של פריטים

נימוק הבחירה - קומפקטיות בזיכרון ושיפור משמעותי בזמן הריצה

מציאת כל הקבוצות התדירות תוך שימוש באלגוריתם FP-Growth, הצגת חוקי ההקשר החזקים, דיווח תוצאת ההרצה, ניתוח והסקת מסקנות.

ההנחה היא ש: Min_Support = 40% , Min_Confidence = 60%

נעבוד עם סט הנתונים המטויב מהפרויקט הקודם. בסט נתונים זה ישנם 1762 רשומות. בכדי שקבוצה תוגדר כשכיחה (Min_supp), כל פריט שמוכל בה צריך להופיע בסט הנתונים לפחות 1762*0.4=705 פעמים.

נביט בקצרה על משתנה המטרה שאותו אנו מעוניינים לחזות:

	Label	Count
1	Insufficient_Weight	231
2	Normal_Weight	233
3	Overweight_Level_I	210
4	Overweight_Level_II	225
5	Obesity_Type_I	268
6	Obesity_Type_II	271
7	Obesity_Type_III	324

כפי שניתן להתרשם, אין אף ערך שחוזר על עצמות 705 פעמים או יותר.

נתמודד עם בעיה זו כפי שהתמודדנו בפרויקט הקודם ע"י דיסקרטיזציה וחלוקה למחלקות (bins) עפ"י מקטעים וקידוד העמודות לערכים בינרים (one-hot-encoding) .

את דיסקרטיזציה נבצע על עמודות שעדיין לא חולקו למחלקות בסט המטוייב, את ההמרה מערכים נומינליים לבינארים נבצע בעזרת Weka כאשר נחלק את משתנה המטרה ל-bins2 כך שהוא מייצג NObeyesdad=CHECK 1/0 (משקל תקין/לא תקין).

פלט התוצאות של ההרצה של האלגוריתם FP-Growth עבור תמיכה מינמלית של 40% וביטחון מינימלי של 60% כלל :

- 100 קבוצות תדירות.
- 44/100 קבוצות מכילות את משתנה המטרה.
- 1/44 מכילה את משתנה המטרה בלבד
- קבוצות התדירות הכוללות את משתנה המטרה מופיעות בקובץ fp-frequent-itemsets.xlsx
- קיימים 332 חוקי הקשר חזקים, ב-43 מתוכם משתנה המטרה מופיע כסיפא .חוקים אלו מופיעים בקובץ fp-association-rules.xlsx

חוקים אלו מלמדים על מידת קשר מסוימת בין צירופי ערכים של מאפיינים שונים לבין סיווג רמת ההשמנה.

קורלציה חיובית/שלילית מראה על הקשר החזק/חלש בין המשתנים .

קורלציה אפסית (lift=0) מראה על אי-תלות, כלומר הצד הגורר לא משפיע על הצד הנגרר.

הסקת מסקנות -

אנו מקבלים חוקי הקשר שמראים על השפעת מצב רמת ההשמנה של אדם בהינתן הרגלי האכילה שלו. חוקי ההקשר החזקים שמצאנו יכולים לתרום להבנה מה עלול לגרום למצב לא תקין.

ניתוח אשכולות – Cluster Analysis

א. נגדיר מדדי איכות לאשכולות

לתהליך ניתוח אשכולות יש מגוון מדדי איכות, בין היתר – יכולת התמודדות עם כמות גדולה של מידע , עם נתונים רועשים, נתונים רב-ממדיים מטיפוסים שונים ויכולת הסתגלות של המודל לנתונים חדשים מבלי להצטרך ללבנות את האשכולות מחדש ועוד.

למדידת האיכות של האשכולות עצמם, דהיינו מדידת איכות של חלוקה מסוימת משתמשים בין היתר במדדי האיכות הבאים –

1. **הומוגניות ושלמות** - עד כמה דומים זה לזה העצמים בכל אשכול ועד כמה הם שונים מעצמים באשכולות אחרים? ככל שגובר הדמיון והמשותף בתוך האשכול ושונה מעצמים מחוץ לאשכול החלוקה יותר איכותית. (בדיקת ה "מרחק" בין עצמים בכל אשכול)

2. **מגמתיות** - האם ישנה איזושהי מגמה או תופעה לא טריוויאלית שניתן ללמוד עליה מהחלוקה? מדד זה יכול להעיד על חלוקה טובה במידה והיא תניב מבנים לא אקראיים שמלמדים אותנו על תובנות משמעותיות שלא ידענו עליהן קודם טרום החלוקה.

ב. בחירת גישה אחת לניתוח אשכולות

מבין הגישות השונות לניתוח אשכולות – בחרתי בשיטת החלוקה באלגוריתם k-means.

בגישות מבוססות חלוקה, מתבצע פיזור של העצמים בין מחיצות זרות כך שהחלוקה מבוססת על פונקציית מרחק/דמיון. בגישות אלו אנו נדרש לציין מראש את k – מספר האשכולות המבוקש, כפרמטר קלט.

אלגוריתם K-Means :

אלגוריתם זה משתמש בעקרון החלוקה ובעזרת עיקרון זה ופונקציית מרחק מתאימה, הוא יוצר k מחיצות בהן הוא מקבץ את n הפריטים מסט הנתונים תוך מזעור ריבועי המרחקים מכל מרכז באשכול. לאחר חלוקה זו, בכל אשכול מחושבת נקודת המרכז (centroid) ביחס לכל הנקודות שנמצאות כעת באשכול ובהתאם לכך מתבצע חישוב מחדש של פונקציית המרחק/דמיון בין כל הנקודות לנקודות הכובד המעודכנות ושוב הנקודות משויכות לנקודת הכובד העדכנית לפי פונקציית המרחק. תהליך זה קורה באופן איטרטיבי עד אשר הקבוצות מתייצבות, נקודת המרכז (centroid) בכל אשכול מתכנות ולא מקובעות או עד אשר מגיעים לתנאי עצירה שהוצב מראש.

הסיבה שבחרתי באלגוריתם זה היא מאחר וDBSCAN טוב יותר לערכים רציפים, במקרה שלנו קיימים ערכים נומינלים / בינארים.

ג. תיאור שלבי ניתוח האשכולות עבור גישת ה K-Means .

1. הכנת הנתונים

רב-ממדיות בנתונים מהווה אתגר לתהליך ניתוח אשכולות. ככל שהממד גבוה יותר נקבל חלוקה "מלוכלכת" יותר, כמות גדולה של מבנים אקראיים ופחות משמעותיים. במקרה שלנו, השתמשתי בסט הנתונים המטויב מממ"ן 21 עליו הפעלתי בבחירת מאפיינים ב- Weka במטרה לצמצם את הממד. כמו כן, ננרמל את המאפיינים הנומריים לטווח שבין 0 ל-1 כדי שפונקציית המרחק/דמיון לא תייחס משקל עודף למשתנים מסוימים על פני אחרים.

סט הנתונים לאחר השינויים לעיל בקובץ : reduced_dataset.csv

2. הפרמטרים וערכיהם

הפרמטרים העיקרים עבור אלגוריתם זה הם ערך ה-k ופונקציית המרחק.

הפרמטר k קובע את מספר האשכולות המבוקשים לחלוקה. בהתאם לפרמטר זה נקבעות k הנקודות הראשונות שיצרו את האשכולות ההתחלתיים.

פונקציית המרחק קובעת את מידת הקירוב והשיוך בהתאם של כל נקודה לאשכול הכי קרוב אליה.

במהלך מספר ניסיונות הרצה בעלי ערכי k שונים, פונקציות מרחק שונות ושיטת התחלה, גיליתי שהערך המינימלי של רשומות שסווגו באופן שגוי מתקבל כאשר k=8 , פונקציית המרחק היא פונקציית מנהטן ושיטת התחלה אקראית.

ד. הרצה ודיווח התוצאות

הפלט של תוצאת ההרצה של k-means מופיע בקובץ :k-means_output.txt

ה. ניתוח התוצאות והסקת מסקנות

קיבלנו שאחוז התוצאות הרשומות שסווגו באופן שגוי הוא 58.9671% כלומר 44.0329% סווגו נכון. לפי חוק ה- base line שמתבסס על רמת השכיחות הגבוהה ביותר, בסט שלנו אנו מקבלים זאת בסיווג של Obesity_Type_III שמופיעה ב- 324 רשומות מתוך 1762 כך שהיא מהווה 18% מהכלל, כך שהדיוק שלנו בשימוש K-Means עובר את "סף המינימום" אך למרות זאת החלוקה לא תורמת יותר מידי.

להלן פילוח הערכים בין האשכולות באלגוריתם K-Means :

Total	0	1	2	3	4	5	6	7	<-- assigned to cluster
231	1	0	2	20	61	89	0	58	Insufficient_Weight
233	20	6	5	24	76	50	4	48	Normal_Weight
210	38	24	0	30	78	27	5	8	Overweight_Level_I
225	49	17	4	44	99	3	6	3	Overweight_Level_II
268	45	42	3	83	93	0	0	2	Obesity_Type_I
271	0	0	0	224	45	1	1	0	Obesity_Type_II
324	0	0	140	0	1	0	183	0	Obesity_Type_III
	153	89	154	425	453	170	199	119	Total

ניתן להבחין בבירור לפי ערכי עמודת המטרה, אילו אשכולות שייכים לכל סיווג.

רשת נוירונים מלאכותית – ANN

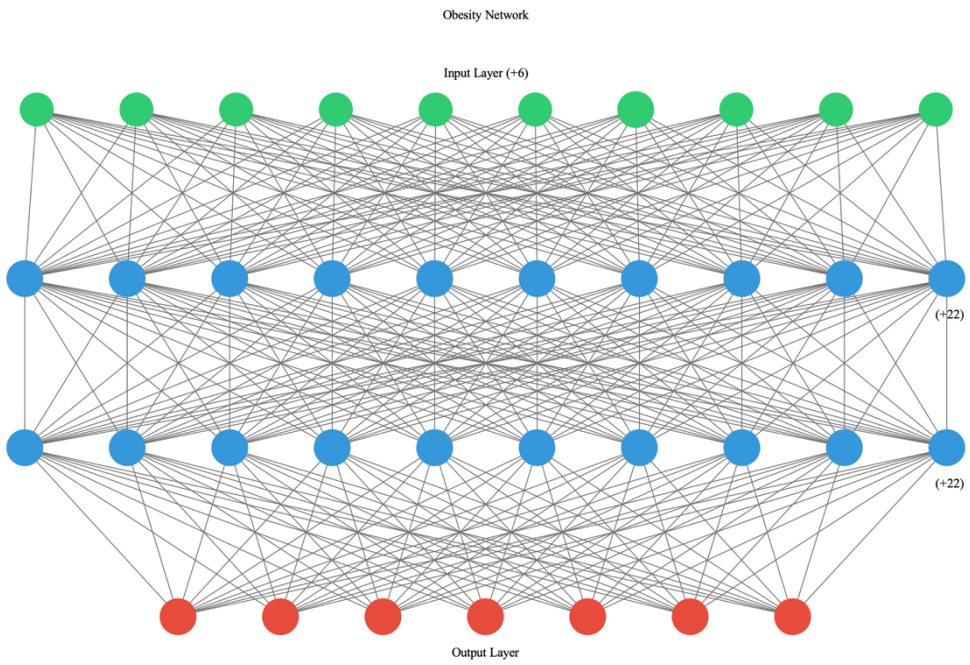
* בחלק זה של הפרויקט, בחרתי להשתמש בסט הנתונים המקורי שנטול ערכים חריגים שהוסרו במטלה הקודמת. (נלקח מקובץ outliers_removed.csv)

א. ארכיטקטורת הרשת

שכבת הקלט – וקטור בעל 16 תכונות.

שכבות חבויות – 2*36 . כל אחת בגודל פי 2 מגודל וקטור הקלט.

שכבת פלט – וקטור בגודל 7 (כמספר הסיווגים האפשריים).



אופן זרימת המידע הוא Feed-forward.

פונקציות האקטיבציה של שכבת הקלט והשכבות החבויות תהיה מסוג "ReLu - Rectified Linear Unit Activation Function" ועבור שכבת הפלט, בחרתי ב "softmax" שנועדה לבעיות סיווג מסוג multiclass .

ב. תהליך האופטימיזציה

בחרתי בפונקציית השגיאה מסוג "categorical_crossentropy" מאחר והיא מתאימה לבעית הסיווג שלנו.

כמו כן, השתמשתי בפונקציית למידה אדפטיבית "Adam" . פונקציה זו תאפשר לנו בחירה אוטומטית של קצב הלמידה תוך כדי הלמידה.

גודל ה batch יהיה 100 – בחרתי לעשות זאת מאחר וחשוב שלא יהיה פער בין רמת הדיוק של קבוצת האימון לבין רמת הדיוק של קבוצת הבדיקה. פער גודל יעיד על התאמת יתר ולכן ערך מצאתי בערך 100 כערך אופטימלי עבור הבעיה.

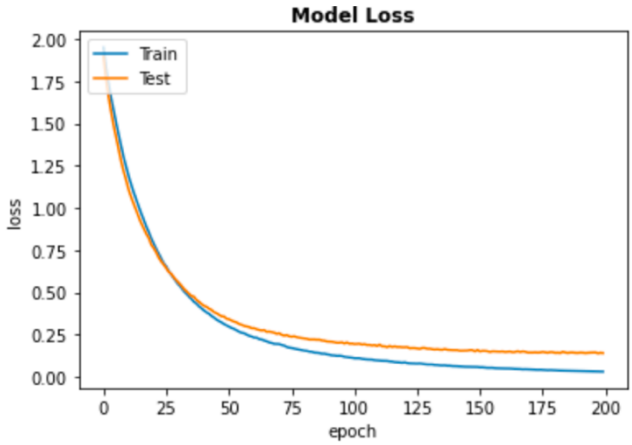
ג. הרצת והערכה של ביצועי הרשת לאורך ה- Epochs של האימון

את ההרצה והתוצאות לאורך כל ה-200 Epochs ניתן לראות בקובץ ANN.ipynb .

אנו מתחילים עם loss: 2.0553 - accuracy: 0.1598 - val_loss: 1.9034 - val_accuracy: 0.1867

ומסיימים עם loss: 0.0296 - accuracy: 1.0000 - val_loss: 0.1686 - val_accuracy: 0.9509

באיור מטה נוכל לראות את גרף השגיאה עבור נתוני האימון ועבור נתוני המבחן כפונקציה של ה- Epochs.



ד. דיווח אודות מקרים בהם בוצע סיווג שגוי

ע"י שימוש בפונקציית predict() מספריית keras, גילינו שבמודל קיימים 27 מקרים חריגים בהם התקבל סיווג שגוי.

ה. ניתוח תוצאות והסקת מסקנות

כחלק מתהליך בניית הרשת, המודל נבדק תחילה עם k-fold על כל סט הנתונים במטרה לחזות את רמת הדיוק שאנו נקבל, והחיזוי היה דיוק של 95.69% .

לאחר כל התהליך של בניית ואימון הרשת,

בהרצת הרשת על כל סט הנתונים, התקבל דיוק של 95.4631%

בהוספת מידע חדש לרשת, קיבלנו דיוק של 95.46%, דיוק קרוב מאוד ומספק במקרה שלנו.

לסיכום, קיבלנו שהרשת נוירונים שבנינו מסווגת בדיוק גבוה את רמת ההשמנה של אותו אדם, כפי שמצופה..

סיכום ומסקנות

התוצאות שקיבלנו בממ"ן זה וגם בממ"ן 21 לא היו שונות אחת מהשניה כי אם בסופו של יום, בממ"ן 21 השתמשנו ביכולת של עצי החלטה לסיווג הבעיה שלנו וכאן ברשת נוירונים מלאכותית. שניהם שימשו אותנו לאותה מטרה ואכן קיבלנו את אותם התוצאות. בפרוייקט זה נחשפנו לשיטות ודרכים למציאת חוקי הקשר, ניתוח אשכולות ובניית רשת נוירונים – כולם כאחד משרתים את אותה המטרה שלשלמה התכנסנו לפני כ- 13 שבועות – כריית מידע. כקורס המשך למבוא לבינה מלאכותית שגרם לי לרצות להעמיק וללמוד את התחום, למדתי המון דברים חדשים, התמודדתי עם בעיות מגוונות, העשרתי את הידע שלי בקריאת מאמרים ונחשפתי לעוד תחומים מעניינים. ובנימה זו, תודה רבה לך על הכל!

נתנאל.