מערכות לומדות וכריית נתונים

פרויקט הקורס – מדד האושר

סמסטר א' שנה"ל תשפ"ג

המחלקה להנדסת תעשייה וניהול, אוניברסיטת בן גוריון

פרופ' בעז לרנר

תאריך הגשה:

23.02.2023

מגישים:

עומרי זאבי 313327041

שרון רבינוביץ 313468175

תקציר

מדד האושר העולמי מחושב באמצעות דו"ח שמתפרסם אחת לשנה על ידי האו"ם, אשר מודד את מידת האושר והרווחה של תושבים במדינות שונות בעולם, על סמך גורמים הקשורים לאיכות חיים. המדד מבוסס על סרק בו המשיבים מתבקשים לדרג את חייהם בסקאלה של 1-10 ביחס לחיים האידיאליים לפי דעתם, בשאלות אשר קשורות למידת העזרה שהם מקבלים מהמדינה בעת צרה, מידת החופש שקיים במדינה שלהם ורמת השחיתות בה.

במהלך בחינת התוצאות של מדד האושר לשנת 2020, ראינו שישראל דורגה במקום מאוד גבוה – 14 מתוך 151 מדינות. דירוג זה העלה לנו את השאלה כיצד ניתן להסביר את הפער בין מציאות החיים כאן שלעתים נראית בלתי אפשרית, מציאות רוויית קונפליקטים, לחצים ומתחים, לבין העובדה שהתושבים בישראל מאושרים יותר מרוב תושבי מדינות העולם.

כפי שצוין לעיל, המדד מתבסס על שאלות סובייקטיביות שנשאלו משיבים ממדינות שונות בעולם – ואילו המטרה שלנו בפרויקט היתה להבין יותר טוב אילו גורמים אובייקטיבים, ברמת המדינה ולא ברמת הפרט, משפיעים על מידת האושר של תושבים במדינה. על כן, בפרויקט זה חקרנו מאפיינים שונים של מדינות ברחבי העולם במטרה להבין כיצד הם משפיעים על ערך מדד האושר של כל מדינה. הפרויקט נועד להעלות את המודעות של ממשלות לגורמים אשר קשורים לאושר ולרווחת התושבים במדינתם ובכך לשפר את איכות חייהם.

לצורך ביצוע הפרויקט, אספנו נתונים אודות מדינות העולם אשר נוגעים בהיבטים שונים, כגון גודל האוכלוסייה, תוחלת חיים, שיעור ההתאבדות ותל"ג לנפש. לאחר הבנת הנתונים, הכנו את הנתונים לקראת המידול בעזרת טיפול בחריגים, השלמת ערכים חסרים, איחוד קטגוריות, נרמול הנתונים ובחירת מאפיינים. בסיום שלב הכנת הנתונים הגענו לסט נתונים אשר כולל 8 מאפיינים ו-148 מדינות. ביצענו אשכול למדינות השונות באמצעות שני אלגוריתמים (אשכול היררכי ואשכול k-means) ובחנו את תוצאותיהם. האשכול הסופי חילק את המדינות לארבעה אשכולות אשר מפרידים בצורה טובה ובעלת משמעות הן את הערכים של המאפיינים של המדינות בכל אשכול והן את ערכי מדד האושר של המדינות בכל אשכול.

בעזרת האשכול והמאפיינים שהובילו כל מדינה להשתייך לכל אשכול, ניתן להסיק מסקנות אודות הגורמים המשפיעים על רווחת התושבים במדינה ולהבין באיזה תחומים יש להשקיע משאבים על מנת להביא לשיפור ערך מדד האושר. להמשך עבודה זו, אנו מציעים לבצע ניתוחי רגישות עבור כל מדינה, אשר יעזרו למדינות שמעוניינות לעלות במדד האושר או לשמור על מדד האושר הקיים שלהן, לפתח אסטרטגיה לשם כך. כלומר, ניתן לבדוק איזה שינוי במאפיינים יכול לגרום למדינה לעבור מהאשכול בו היא נמצאת לאשכול אחר.

תוכן עניינים

מבוא והבנת התחום

איכות חיים מוגדרת על ידי ארגון הבריאות העולמי כ"תפיסת הפרט לגבי מיקומו בחיים בהקשר של התרבות ומערכות הערכים שבהן הוא חי וביחס למטרותיו, הציפיות, הסטנדרטים והדאגות שלו". כשמבצעים הערכה של איכות החיים, יש להעריך את מצב הרווחה של אדם, קבוצת אנשים או רווחתם של אזור או אומה. איכות חיים טובה של התושבים במדינה היא משמעותית הן עבור התושבים עצמם והן עבור מקבלי ההחלטות בממשלה, שהרי ממשלה שפוגעת באספקט מסוים באיכות החיים של תושביה עלולה להיתקל במחאות. מדינות שונות תופסות את איכות החיים של תושביהן באופן שונה, אך רוב המדינות מספקות מגוון שירותים לרווחת התושבים במגוון תחומי חיים, על מנת שיהיו שבעי רצון מהמגורים במדינה.

בין איכות החיים לבין אושר ישנו קשר הדוק. בספרות, אושר מוגדר לעתים קרובות כרווחה סובייקטיבית, רווחה רגשית, השפעה חיובית ואיכות חיים (Diener, 2000). במקומות אחרים, אושר סובייקטיבי הוגדר כ"הערכה עולמית של שביעות רצון מהחיים" (Diener, 2006). הגדרות אלו מצביעות על קשר מובהק בין המבנים של אושר, רווחה סובייקטיבית, איכות חיים ושביעות רצון מהחיים. ניתן אף להניח כי משמעויות האושר השונות עשויות להיות תלויות בהקשר ושמושג האושר נתון לפרשנויות שונות המוצעות על ידי אנשים שונים (Carlquist et al., 2016).

מאז שנת 2012, כמעט בכל שנה מתפרסמות תוצאות מדד האושר העולמי כחלק מדו"ח האושר של האו"ם, המדרג את מדינות העולם לפי מידת האושר של התושבים בהן. מדד האושר הוא בעצם סקר מקיף שמעריך אושר, רווחה והיבטים של קיימות וחוסן בקרב תושבי המדינה. סקר זה יכול לשמש למדידת שביעות רצון מהחיים, ובנוסף יכול לשקף את אי השוויון בהכנסה, מידת האמון בממשלה, תחושת קהילתיות והיבטים אחרים של רווחה בקרב אוכלוסייה ספציפית (Musikanski et al., 2017). זהו כלי יחיד מסוגו הזמין באופן חופשי ברחבי העולם. בדרך כלל תוצאות הסקר מובילות לדיונים בתקשורת בנוגע לשאלה האם אנחנו מאושרים יותר או פחות מהשנה שעברה (Carlsen, 2020). המדד מבוסס על שבעה אינדיקטורים אשר מסוכמים לערך המדד, המשמש בסופו של דבר לדירוג של יותר מ-150 מדינות לפי האושר הנתפס של התושבים בהן. כל המשתתפים בסקר צריכים לדרג בציון של 0 עד 10 (לפי סולם Cantril) שאלות אשר מתבטאות בשישה גורמים המשפיעים על ציון האושר: תוצר לאומי גולמי, תוחלת חיים, נדיבות, תמיכה חברתית, חופש ושחיתות.

על אף האמור לעיל, ישנן ביקורות המופנות כלפי השימוש במדד האושר העולמי בצורתו הנוכחית. המבקרים מציינים כי יש הבדל בין חוויה של רווחה לבין הערכה שלה (Kushlev et al., 2015). לדוגמה, קולומביה הגיעה למקום ה-37 בדירוג האושר העולמי בשנת 2018, אך הגיעה למקום הראשון במדד החוויה החיובית של Gallup. בנוסף, ישנו חוסר בעקביות של תוצאות של סקרי מדידת אושר שונים. למשל, סקר Pew שנערך ב-43 מדינות בשנת 2014 (שלא כלל את רוב אירופה) גרם למקסיקו, ישראל ו-ונצואלה לסיים במקום הראשון, השני והשלישי בהתאמה. מבקרים אחרים מציינים כי המשתנים בהם משתמשים בדו"ח האושר העולמי מתאימים יותר למדידת אושר ברמה הלאומית ולא ברמת הפרט.

הספרות בנושא חיזוי מידת האושר באמצעות מערכות לומדות היא דלה יחסית. בין העוסקים בנושא, ניתן לציין את מחקרם של Jannani et al., 2021 אשר ניסו לחזות את ציון מדד האושר העולמי לשנת 2021 באמצעות אלגוריתמים שונים של רגרסיה. בכלים בהם השתמשו, הצליחו להראות כי רגרסיה ליניארית מרובה, רגרסיית לאסו ו-LSTM היו האלגוריתמים שנתנו את הביצועים הטובים ביותר עבור בעיה זו.

הפרויקט שלנו מתעסק באשכול המדינות השונות הנכללות בדו"ח האושר העולמי לשנת 2020 בהתאם למאפיינים רבים, נוסף על אלו הקיימים בדו"ח, כדי לבחון את השפעתם על ערך מדד האושר. נרצה לבחון כיצד ערכי המאפיינים שבחרנו משפיעים על השיוך של מדינה מסוימת לאשכול מסוים. בהתאם לנלמד בקורס, נעבוד לפי עקרונות ה-CRISP-DM. בשלב מידול הבעיה, המגיע לאחר הבנת הנתונים והכנתם, נבצע חלוקה לקבוצות באמצעות אלגוריתמי אשכול, ולבסוף נעריך את תוצאות המודלים שהתקבלו אל מול ערך מדד האושר של המדינות. באמצעות בחינת התוצאות שיתקבלו, ניתן יהיה להפיק תועלת בכמה דרכים. למשל, ממשלות יכולות לבצע רפורמות ברמה המדינית על מנת להגדיל את מידת האושר שהתושבים חשים ואף להעריך את האפקטיביות של תוכניות קיימות. על כן, פרויקט מסוג זה עשוי לסייע למדינות להבין את הגורמים העיקריים אשר תורמים למדד האושר ובכך לקבל החלטות טובות יותר (Jannani et al., 2021).

הבנת הנתונים

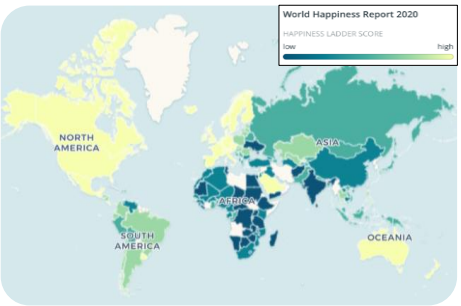
תיעוד מקור הנתונים

יצרנו את בסיס הנתונים באמצעות איחוד נתונים אשר הפקנו מחמישה מקורות שונים: מדד האושר (World Happiness Report), אתר הבנק העולמי (The World Bank), מחלקת הסטטיסטיקה של האו"ם (The United Nations Statistics Division), דו"ח האוכלוסין העולמי של לשכת האוכלוסין האמריקאית (World Population Review) והיחידה לחקר האקלים (Climatic Research Unit).

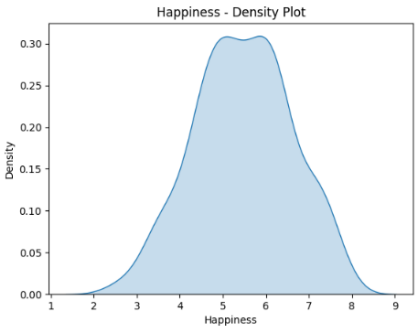
בסיס הנתונים מכיל נתונים על 151 מדינות משנת 2020. הנתונים כוללים 15 מאפיינים – 14 רציפים ואחד קטגוריאלי, אשר הערכנו בתחילת ביצוע הפרויקט כי יש להם קשר לרמת האושר של התושבים במדינה. משתנה המטרה בבסיס הנתונים הוא ציון מדד האושר שקיבלה כל מדינה בשנת 2020.

בפרויקט זה נבצע אשכול של המדינות על פי הגורמים השונים המאפיינים אותן. לאחר מכן, נבחן וננתח את תוצאות האשכול אל מול ערך מדד האושר של המדינות בכל אשכול, במטרה לזהות קשר בין מאפייני האשכול לבין מדד האושר של המדינות שנמצאות בו.

פירוט המאפיינים בבסיס הנתונים מופיע בנספח 1. המשתנה Country לא ייכלל בניתוחינו משום שהוא לא נועד להסביר את משתנה המטרה או להשפיע על האשכול, אלא רק מהווה שדה ייחודי לכל רשומה.



איור 1 - ערך מדד האושר של המדינות ברחבי העולם מתוך דו"ח האושר העולמי של האו"ם לשנת 2020

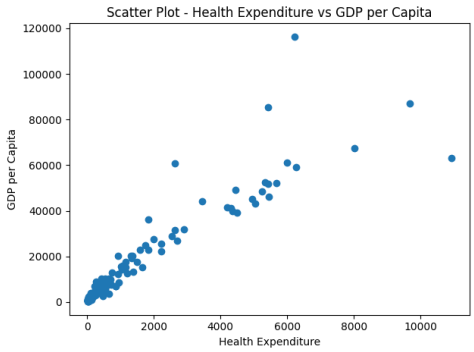


איור 2 - צפיפות משתנה המטרה

באיור 1 ניתן לראות את ערך מדד האושר של המדינות ברחבי העולם על גבי מפת העולם. באיור 2 מוצג גרף הצפיפות של משתנה המטרה, ערך מדד האושר, אשר מלמד על ההסתברות האפריורית שלו. ניתן לראות כי המשתנה מתפלג נורמלית סביב הערך 5.5. ערך מדד האושר ניתן על פי סולם Cantril שנע בין 0-10, כאשר המשיבים התבקשו לדרג את החיים המיטביים עבורם בציון 10 ואת החיים הכי גרועים עבורם בציון 0, ולאחר מכן לדרג את חייהם הנוכחיים על פי אותו סולם. על כן, עקומת הפעמון שהתקבלה תואמת להגיון – מרבית המדינות קיבלו ערך מדד אושר בינוני, בין 4-7, וככל שמתקרבים לקצוות של ערך המדד, מספר המדינות פוחת.

סטטיסטיקה תיאורית

ניתוח ראשוני של המשתנים הרציפים, הכולל ממוצע, חציון וערכי אחוזונים מופיע בנספח 2. תרשימי צפיפות עבור המשתנים הרציפים ותרשים היסטוגרמה עבור המשתנה הקטגוריאלי מופיעים בנספח 3. בבחינת המשתנים המסבירים הרציפים, ניתן לראות כי מרביתם חסומים עם גבול תחתון באפס ובעלי זנב ימני הכולל מעט מדינות בעלות ערכים גבוהים מאוד. באיור 3 ניתן להבחין בקורלציה גבוהה בין התל"ג לנפש לבין הוצאות המדינה על בריאות לנפש, והדבר הגיוני מכיוון שככל שמדינה עשירה יותר, כך היא יכולה להרשות לעצמה להשקיע סכומי כסף גדולים יותר באופן כללי ובפרט בתחום הבריאות.

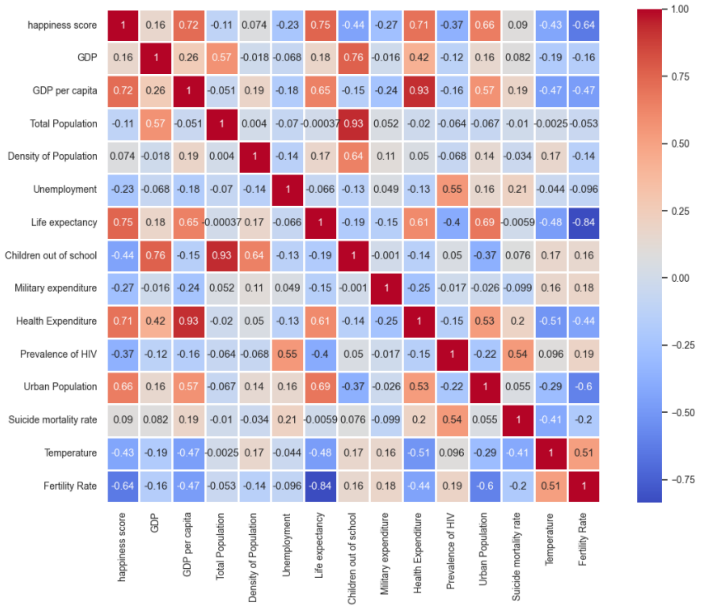


איור 3 - הוצאות בריאות מול תל"ג לנפש

קשרים בין משתנים

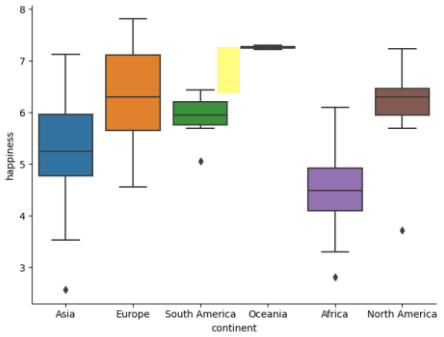
נבצע השוואה בין המשתנים הרציפים על פי מקדם המתאם של פירסון, כמדד לקשר ליניארי. התוצאות מוצגות בתרשים Heatmap באיור 4. בנוסף, נבחן את הקשרים בין המשתנים באמצעות תרשים Pair plot הכולל תרשימי פיזור בין המשתנים הרציפים (נספח 4).

ניתן לראות כי המשתנה המוסבר נמצא בקורלציה חיובית גבוהה עם מספר משתנים מסבירים כגון תל"ג לנפש, תוחלת חיים, הוצאות המדינה על הבריאות, ובקורלציה נמוכה עם המשתנים שיעור התאבדות וצפיפות האוכלוסייה. סביר להניח כי שיעורי ההתאבדות מייצגים תופעה המתרחשת בקנה מידה קטן יחסית, כך שהיא לא מעיבה על תחושת האושר של התושבים.



איור 4 - תרשים Heatmap

באיור 5 ניתן לראות את התפלגות ציוני מדד האושר בהתאם לכל יבשת. נראה כי באירופה באופן כללי ציוני מדד האושר גבוהים יותר משאר היבשות, וביבשת אפריקה הם הנמוכים ביותר. כמו כן, יבשת אוקיאניה מכילה רק שתי מדינות – אוסטרליה וניו זילנד, ששתיהן בעלות ציוני אושר גבוהים במיוחד. בהמשך נבצע איחוד קטגוריות במשתנה היבשת, על מנת לתת תמונה מייצגת יותר של התפלגות האושר ביבשות השונות.



איור 5 - ערך מדד האושר על פי חלוקה ליבשות

הכנת הנתונים

המשתנים תל"ג, תל"ג לנפש ואוכלוסייה מוכלים אחד בתוך שני (משתנה תל"ג הוא תוצאת המכפלה של שני המשתנים האחרים), לכן נסיר את אחד מהם בשביל להימנע מיתירות. בנוסף, באיור 4 ניתן לראות מולטיקולינאריות גבוהה מאוד בין תל"ג לנפש לבין הוצאות בריאות (0.93) וגם זו בעיה שיש לטפל בה. על מנת להתמודד עם שתי הבעיות לעיל, החלטנו להסיר את משתנה תל"ג לנפש. למשתנה זה יש קורלציה גבוהה עם המשתנה המוסבר (0.72) אך אנו מקבלים אותה גם ממשתנה הוצאות בריאות (0.71).

לאחר מכן, מאחר והתחלנו עם יחסית הרבה משתנים מסבירים ביחס לכמות התצפיות, רצינו להסיר עוד כמה מהם על מנת להקטין את מימד הבעיה. כך נקטין את הסיבוכיות, נטפל בקללת המימד ונוכל להגיע לביצועים טובים יותר. בחרנו להסיר את המשתנים עם הקורלציה הנמוכה ביותר אל מול המשתנה המוסבר: התאבדות, צפיפות, אבטלה ואוכלוסייה (|0.23|>).

טיפול בחריגים

באופן כללי, הגישה שלנו היא לא להסיר תצפיות מכיוון שאנחנו מראש מתחילים עם יחסית מעט מדינות. הפקנו תרשימי קופסה לכל המשתנים הרציפים ובדקנו את הערכים החריגים בכל משתנה. כל התצפיות החריגות מייצגות את המציאות ולא מהוות טעות בנתונים, לכן בחרנו להשאיר אותן ולהתחשב בהן כחלק ממשימת האשכול. למשל, במשתנה גודל האוכלוסייה יש כמה ערכים חריגים גבוהים במיוחד, אך הם שייכים למדינות גדולות מאוד כגון סין והודו, אותן אנו מעוניינים להשאיר בניתוח הנתונים.

טיפול בערכים חסרים

לפי טבלה 1, במשתנה אחוז הילדים מחוץ למסגרת חינוכית יש מעל 70% ערכים חסרים ולכן בחרנו להסיר אותו. בשלב זה ביצענו חיפוש במקורות נוספים על מנת להשלים את הערכים החסרים. במהלך החיפוש זיהינו שלוש מדינות עם הרבה ערכים חסרים שלא הצלחנו להשלים: קוסובו, עזה והונג קונג. למרות שהגישה שלנו היתה לא להסיר תצפיות, החלטנו להסיר את שלושת המדינות האלה מכיוון ששלושתן לא מוגדרות כמדינות רשמיות. בסיום התהליך הגענו למצב הערכים החסרים המתואר בטבלה 2.

Variable	# Missing values	% Missing values
Happiness Score	0	0%
GDP	0	0%
Life expectancy	0	0%
Military expenditure	11	7.3%
Health Expenditure	0	0%
Prevalence of HIV	13	8.6%
Urban Population	0	0%
Continent	0	0%
Temperature	0	0%
Fertility Rate	0	0%

טבלה 1 - פירוט ערכים חסרים בסט הנתונים לאחר השלמה ידנית

Variable	# Missing values	% Missing values
Happiness Score	0	0%
GDP	3	2%
Life expectancy	0	0%
Children Out of School	107	70.9%
Military expenditure	14	9.3%
Health Expenditure	6	4%
Prevalence of HIV	33	21.9%
Urban Population	1	0.7%
Continent	0	0%
Temperature	3	2%
Fertility Rate	18	11.9%

טבלה 2 - פירוט ערכים חסרים בסט הנתונים לפני השלמה ידנית

בשלב זה נותרו 2 משתנים עם ערכים חסרים. את המשתנה הוצאות צבאיות של הממשלה בחרנו להסיר בעקבות הקורלציה הדי נמוכה שלו עם המשתנה המוסבר (-0.27). את המשתנה שכיחות HIV באוכלוסייה לא רצינו להסיר ולכן בחרנו להשלים אותו בעזרת משתנים אחרים בדאטה. זיהינו קורלציה יחסית גבוהה בינו לבין המשתנים אבטלה ושיעור התאבדות, לכן הרצנו מודל רגרסיה לינארית שחוזה את HIV באמצעותם (נספח 5). במודל שהתקבל שני המשתנים יצאו מובהקים, לכן השלמנו על פיו את הערכים החסרים:

$$\overline{HIV} = -2.818 + 0.313 \cdot uneployment + 0.22 \cdot suicide$$

איחוד קטגוריות

ביצענו איחוד קטגוריות עבור המשתנה הקטגוריאלי יבשת, משום שבחלק מהקטגוריות שלו היו מעט תצפיות ורצינו להימנע מריבוי משתני דמה שלא לצורך. בטבלה 3 ניתן לראות את מספר המדינות בכל יבשת לפני ואחרי איחוד הקטגוריות. בסיום שלב הכנת הנתונים נשארנו עם 8 משתנים מסבירים ו-148 מדינות.

Continent	# Before merge	# After merge
Africa	44	44
Asia	40	42
Oceania	2	
Europe	39	39
North America	13	23
South America	10	

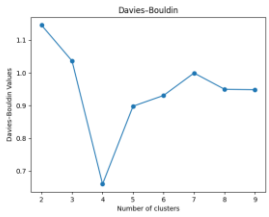
טבלה 3 - איחוד קטגוריות של המשתנה 'יבשת'

נרמול הנתונים

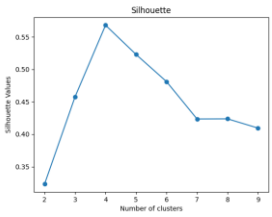
ביצענו נרמול לכלל הנתונים. ראשית ביצענו נרמול באמצעות Robust scale מאחר והוא שימושי כאשר ישנן חריגות בקצוות כפי שתיארנו שמתקיים בנתונים שלנו (לדוגמה למשתנה תל"ג יש טווח מאוד גדול שעלול להשפיע על הנרמול). לאחר מכן ביצענו נרמול Min Max על מנת שכל הערכים ינועו בין 0 ל-1.

מידול

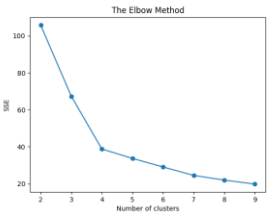
ראשית הרצנו מספר מדדים לבחינת מספר האשכולות המיטבי.



איור 6 - Davies Bouldin



איור 7 - Silhouette



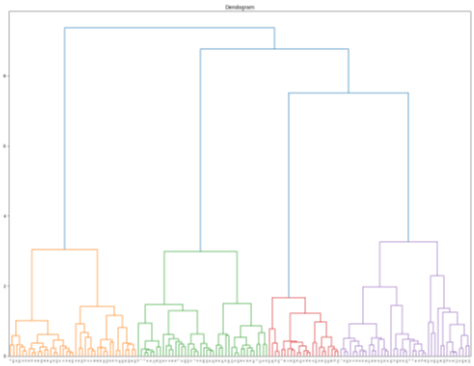
איור 8 - SSE

לפי Davies-Bouldin, הבוחן את יחס המרחקים בתוך האשכולות ובין האשכולות, נרצה 4 אשכולות למזעור המדד. לפי SSE, נרצה למזער את השגיאה ולכן גם נבחר 4 אשכולות. לבסוף, לפי מדד Silhouette, הבוחן מרחקים בין כל תצפית לבין הסנטרואיד של האשכול, נרצה למקסם ונבחר גם כן 4. לסיכום, לפי כל המדדים נראה כי כדאי לאשכל עם 4 אשכולות. כעת נבצע אשכול היררכי ואשכול לפי K-means, נבחן את האשכולות השונים ואת משתנה המטרה, מדד האושר, בכל אשכול.

אשכול היררכי

באיור 9 ניתן לראות את תרשים הדנדוגרם של סט הנתונים אשר מראה את השונות בתוך האשכולות עבור כל רמת אשכול היררכי. ניתן לראות שגם הדנדוגרם תומך בחלוקה ל-4 אשכולות, כאשר איחוד נוסף של האשכולות מ-4 ל-3 גורם לעלייה חדה בשונות בתוך האשכולות.

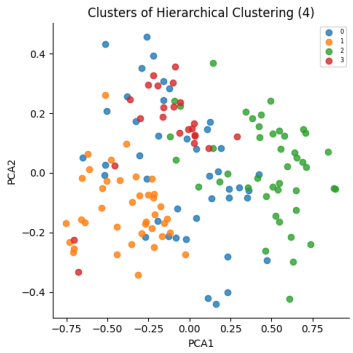
כעת הרצנו PCA על המסבירים לצורך הצגה ויזואלית של האשכולות. PCA1 מסביר כ-60.3% מהשונות ו-PCA2 מסביר כ-13.1% (סך הכל 73.4% שונות מוסברת). באיור 11 ניתן לראות די בבירור הפרדה בין ארבעת האשכולות. יש חפיפה של האשכול האדום עם הכחול, ומעט תצפיות חריגות של האשכול האדום שנמצאות בתחום של האשכול הכתום.



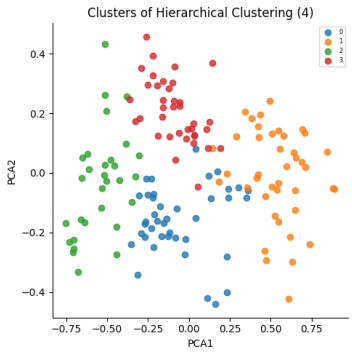
איור 9 - תרשים דנדוגרם

בטבלה 4 מתואר הפילוג של כל אשכול ליבשות. ניתן לראות כי האשכול נעשה באופן מוחלט בהתאם ליבשת אליה שייכת כל מדינה, מה שהעלה חשש כי משתנה זה השפיע יתר על המידה על תוצאת האשכול. לכן, על מנת לוודא שהאשכול שהתקבל לא נוצר לחלוטין כתוצאה ממשתנה יבשת, ביצענו את האשכול בשנית, ללא משתנה יבשת. באיור 11 ניתן לראות את האשכול החדש לפי ה-PCA לשני רכיבים.

ניתן לראות שהסרת המשתנה יבשת הביא לתוצאות אפילו יותר טובות – כעת ההפרדה בין האשכולות האדום והכחול הרבה יותר ברורה, ואין את התצפיות החריגות של האשכול האדום שהופיעו קודם. לפי טבלה 5, עדיין בכל אשכול יש יבשת דומיננטית (עמודת top), אבל כעת החלוקה לא נעשתה באופן מוחלט לפי היבשת (עמודת unique), ובנוסף חלוקת המדינות לאשכולות התאזנה – 37-41 מדינות בכל אשכול במקום 23-44 בכל אשכול (עמודת count). אנו מאמינים שהאשכול ללא משתנה יבשת מתייחס בצורה טובה יותר **למאפיינים של המדינות ולא רק ליבשת בה הן נמצאות**. כלומר, סביר שמדינה שאינה נמצאת ביבשת אפריקה, אבל דומה במאפייניה למדינות אפריקה, תשויך לאשכול השני ולא לאשכול של היבשת שלה. באותו אופן, מדינה מערבית שנמצאת במזרח תשויך לאשכול הרביעי. לכן בחרנו להמשיך ללא המשתנה יבשת.



איור 10 - PCA עבור 4 אשכולות – עם משתנה יבשת



איור 11 - PCA עבור 4 אשכולות – בלי משתנה יבשת

Cluster	count	unique	top	freq
0	42	1	Asia	42
1	44	1	Africa	44
2	39	1	Europe	39
3	23	1	America	23

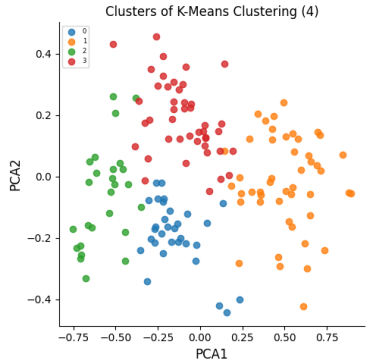
טבלה 4 - פילוג האשכולות לפי יבשת – עם משתנה יבשת

Cluster	count	unique	top	freq
0	40	2	Asia	21
1	41	3	Africa	38
2	30	3	Europe	20
3	37	3	America	19

טבלה 5 - פילוג האשכולות לפי יבשת – בלי משתנה יבשת

אשכול k-means

בשלב זה רצינו לבחון אלגוריתם אשכול נוסף ולכן ביצענו אשכול K-means, עם 4 אשכולות, ללא משתנה יבשת. באיור 12 ניתן לראות שהתקבלו תוצאות דומות לאלו שהתקבלו על פי האשכול ההיררכי. שוב ניכרת הפרדה ברורה בין ארבעת האשכולות, ובמרכז הגרף, בחיבור בין האשכולות, נראה שההפרדה אף טובה יותר מאשר באשכול ההיררכי. בטבלה 6 מתואר הפילוג ליבשות של האשכולות שנוצרו לפי אשכול k-means.



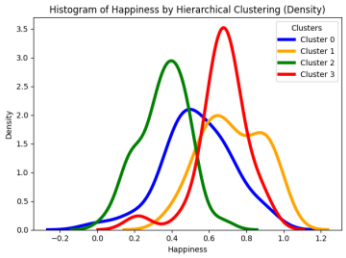
איור 12 - PCA עבור 4 אשכולות – בלי משתנה יבשת

Cluster	count	unique	top	freq
0	31	2	Europe	19
1	47	3	Africa	38
2	26	3	Europe	18
3	44	4	America	19

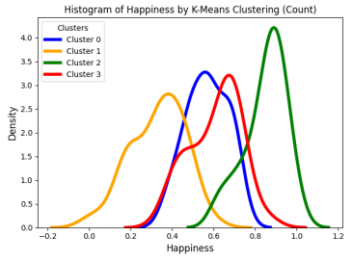
טבלה 6 - פילוג האשכולות לפי יבשת - אשכול K-means

סיכום

להשוואה בין שני אלגוריתמי האשכול, בחנו את ההפרדה של כל אחד מהם בצורה ויזואלית על פי ההתפלגות של המשתנה המוסבר בכל אחד מארבעת האשכולות (איורים 13 ו-14). ניתן לראות שאלגוריתם k-means הפריד בצורה טובה יותר את האשכולות הירוק והכתום אחד מהשני ומהאשכול הכחול. עם זאת, נראה שהוא מפריד פחות טוב את האשכול הכחול מהאשכול האדום מאשר האשכול ההיררכי.



איור 13 - גרף צפיפות - אשכול היררכי



איור 14 - גרף צפיפות - אשכול k-means

על מנת להכריע בין האלגוריתמים, בדקנו את טיב האשכול לפי שלושה מדדים כמתואר בטבלה 7. את מדד silhouette נרצה למקסם ולכן על פיו נעדיף את אשכול k-means. את מדד calinski harabasz נרצה למקסם ולכן על פיו נעדיף את האשכול ההיררכי. את מדד davies bouldin נרצה למזער ולכן על פיו נעדיף את אשכול k-means. לסיכום, **נבחר להמשיך עם אשכול k-means**, בעל ביצועים טובים יותר.

index	Hierarchical	k-means
silhouette score	0.620	0.623
calinski harabasz score	369.409	362.955
davies bouldin score	0.616	0.562

טבלה 7 - מדדים לבחינת טיב האשכול

נתבונן בטבלה 8 אשר מציגה את מרכזי כל האשכולות (סנטרואידים) של אשכול k-means, כאשר הערך המקסימלי עבור כל משתנה (בכל עמודה) צבוע בירוק והמינימלי באדום. כבר ניתן לראות שהתקבלו שני אשכולות קיצוניים (1 ו-2) אשר כמעט כל הערכים הקיצוניים נמצאים בשורות שלהם, וכי שני האשכולות האחרים (0 ו-3) הם בעלי ערכים בינוניים.

Cluster	Happiness Score	GDP	Life Expectancy	Health Expenditure	Prevalence of HIV	Urban Population	Temp	Fertility Rate
0	5.55	6.15E+11	74.45	633.19	0.61	58.87	7.78	1.88
1	4.29	1.10E+11	63.71	88.93	4.19	38.90	23.70	4.17
2	7.02	1.88E+12	81.84	4958.60	0.39	83.47	8.29	1.63
3	5.75	2.26E+11	75.90	706.35	0.57	71.63	22.68	2.15

טבלה 8 - סנטרואידים לפי אשכול kmeans

הערכה

איור 15 - תרשים רדאר לפי אשכול k-means

נתבונן במשמעות האשכולות השונים שקיבלנו על פי אשכול k-means ל-4 אשכולות, כפי שמופיע באיור 15:

אשכול 1 (אדום) – מדינות אלו הן בעלות הערכים הנמוכים ביותר במדדים להלן: תל"ג, תוחלת חיים, הוצאות בריאות של הממשלה לנפש, % תושבים שחיים באזור עירוני. כמו כן הן בעלות הערכים הגבוהים ביותר במדדים להלן: שכיחות HIV באוכלוסייה, טמפרטורה ושיעור ילודה. מרבית המדינות מיבשת אפריקה. המדינות באשכול זה שאינן מיבשת אפריקה הן מדינות שדומות במאפיינים שלהן לאפריקה כגון: אפגניסטן, האיטי, הודו, קמבודיה, לאוס, סרי לנקה, מיאנמר, פקיסטן. מדינות באשכול זה הן בעלות **ערך מדד האושר הנמוך ביותר**.

אשכול 2 (ירוק) – מדינות אלו הן בעלות הערכים הגבוהים ביותר במדדים להלן: תל"ג, תוחלת חיים, הוצאות בריאות של הממשלה לנפש, % תושבים שחיים באזור עירוני. כמו כן הן בעלות הערכים הנמוכים ביותר במדדים להלן: שכיחות HIV באוכלוסייה, טמפרטורה ושיעור ילודה (הפוך מאשכול 1). מרבית המדינות הן ממערב וצפון אירופה (כגון גרמניה, נורבגיה, שוודיה, צרפת, פינלנד).
מדינות נוספות שאינן מיבשת אירופה הן מדינות שנחשבות למערביות ומפותחות: אוסטרליה, קנדה, צ'ילה, ישראל, יפן, ניו זילנד וארה"ב. מדינות באשכול זה הן בעלות **ערך מדד האושר הגבוה ביותר**.

שני האשכולות הנותרים הם בעלי ערכים בינוניים במסבירים וכן במשתנה המוסבר, ביחס לאשכולות 1 ו-2.

אשכול 0 (כחול) + אשכול 3 (סגול) – המדינות באשכול 3 הן בעלות ערכים גבוהים יותר בתוחלת החיים ובהוצאות בריאות של הממשלה לנפש ביחס למדינות באשכול 0. לעומת זאת, המדינות באשכול 0 הן בעלות תל"ג גבוה יותר ושכיחות HIV באוכלוסייה נמוכה יותר ביחס למדינות באשכול 3. מרבית המדינות מאשכול 0 הן ממזרח אירופה, כגון ליטא, בלרוס, סרביה, לטביה, סלובקיה, סלובניה, אוקראינה (מדינות ברית המועצות לשעבר). מדינות נוספות שאינן מיבשת אירופה הן מיבשת אסיה: גאורגיה, נפאל, טורקיה. לעומת זאת, מרבית המדינות מאשכול 3 הן מדרום ומרכז אמריקה, כגון: ארגנטינה, בוליביה, קולומביה, מקסיקו (מדינות חמות בעיקר). מרבית המדינות הנוספות באשכול זה הן מדינות ערב: עיראק, ירדן, לבנון, לוב, מרוקו, ערב הסעודית, איחוד האמירויות ועוד. המדינות באשכול 3 קיבלו ערך מדד אושר גבוה יותר מאשר המדינות באשכול 0.

סיכום, דיון ומסקנות

לסיכום, אנו מאמינים שעבודה זו יכולה לשמש מדינות על מנת להעלות את ערך מדד האישור שלהן ובכך לשפר את איכות החיים של תושביהן. לאור התוצאות שהתקבלו, נראה כי האשכול אכן מתאים לנתונים והביא לחלוקה הגיונית בין המדינות השונות. מצאנו כי חלוקה זו מפרידה בצורה טובה גם את ערכי מדד האושר של המדינות בכל אשכול. באמצעות כך, ניתן לזהות אלו מאפיינים משפיעים על ערך מדד האושר ובאיזה אופן.

בעזרת אשכול המדינות והמאפיינים שהובילו כל מדינה להשתייך לכל אשכול, ניתן להסיק מסקנות אודות הגורמים המשפיעים על רווחת התושבים במדינה ולהבין באיזה תחומים יש להשקיע משאבים שיביאו לשיפור ערך מדד האושר.
להמשך עבודה זו, אנו מציעים לבצע ניתוחי רגישות עבור כל מדינה, אשר יעזרו למדינות שמעוניינות לעלות במדד האושר או לשמור על מדד האושר הקיים שלהן, לפתח אסטרטגיה לשם כך. כלומר, ניתן לבדוק איזה שינוי במאפיינים יכול לגרום למדינה לעבור מהאשכול בו היא נמצאת לאשכול אחר. לדוגמה: איזה שינויים צריכה מדינה ממזרח אירופה לעשות בשביל לעבור לאשכול של המדינות ממערב אירופה, שמקבלות ערך מדד אושר גבוה יותר? עד כמה מדינה שנמצאת באשכול עם מדד האושר הגבוה ביותר, רגישה לשינויים קטנים במאפיינים שלה ועלולה לעבור לאשכול עם מדד אושר נמוך יותר?

למחקר נוסף אנו ממליצים לחקור מאפיינים נוספים אשר חשודים כמשפיעים על ערך מדד האושר על מנת לקבל תמונה רחבה יותר. בנוסף, ניתן להשתמש באשכול שנעשה בעבודה זו לבחינת משתני מטרה נוספים מלבד מדד האושר ולחקור את האשכולות על פיהם, כמו לדוגמה תל"ג או תוחלת חיים.

ביבליוגרפיה

Carlsen, L. (2020). How happy are we actually? A posetic analysis of the world happiness index 2016–2019 Denmark as an exemplary case. International Journal of Community Well-Being, 3(3), 311-322.

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. American psychologist, 55(1), 34.

Diener, E. (2006, November). Guidelines for national indicators of subjective well-being and ill-being. In Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being. Springer.

Jannani, A., Sael, N., & Benabbou, F. (2021, December). Predicting Quality of Life using Machine Learning: case of World Happiness Index. In 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT) (pp. 1-6). IEEE.

Kushlev, K., Dunn, E. W., & Lucas, R. E. (2015). Higher income is associated with less daily sadness but not more daily happiness. Social Psychological and Personality Science, 6(5), 483-489.

Musikanski, L., Cloutier, S., Bejarano, E., Briggs, D., Colbert, J., Strasser, G., & Russell, S. (2017). Happiness index methodology. Journal of Social Change, 9(1), 2.

נספחים

נספח 1 – פירוט המאפיינים

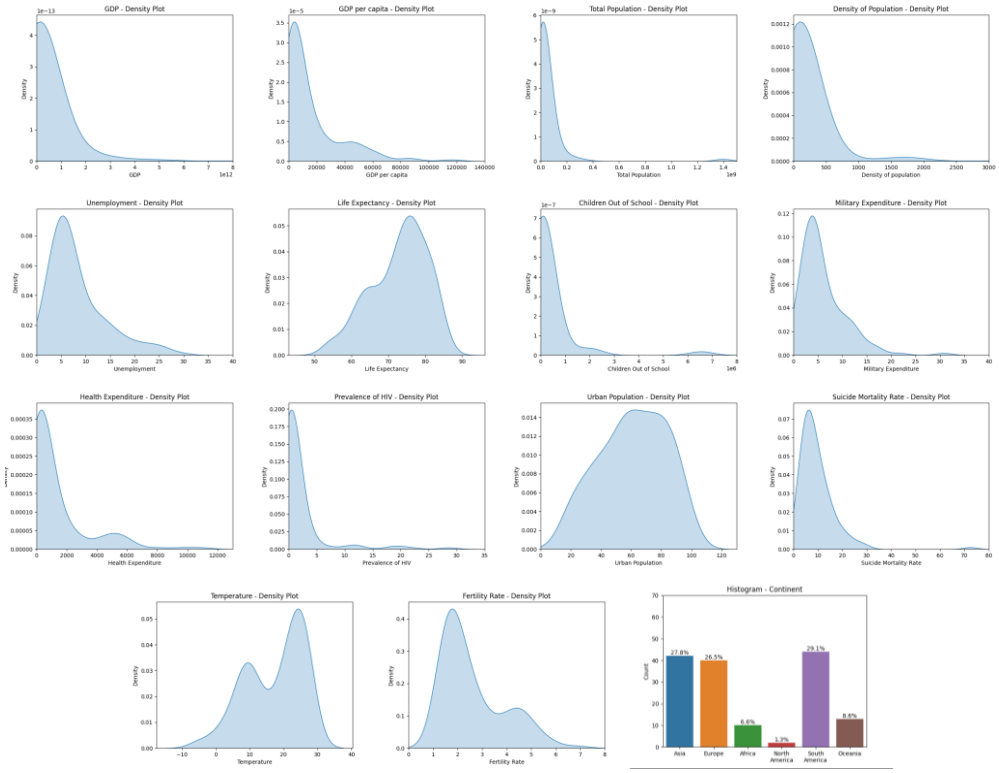
מקור	סוג	פירוט	מאפיין
World Happiness Report	רציף	מדד האושר	**Happiness Score**
World Bank Open Data	רציף	תל"ג ($US)	**GDP**
	רציף	תל"ג לנפש ($US)	**GDP per Capita**
	רציף	סך האוכלוסייה במדינה	**Total Population**
	רציף	צפיפות האוכלוסייה (מספר תושבים לק"מ מרובע)	**Density of Population**
	רציף	אבטלה (% מתוך סך כוח העבודה)	**Unemployment**
	רציף	משך שנות חיים ממוצע של אדם במדינה (שנים)	**Life Expectancy**
	רציף	ילדים מחוץ למסגרת החינוך	**Children Out of School**
	רציף	הוצאות צבאיות (% מהוצאות הממשלה)	**Military Expenditure**
	רציף	הוצאות בריאות של הממשלה לנפש ($US)	**Health Expenditure**
	רציף	שכיחות HIV באוכלוסייה (% מהאוכלוסייה בגילים 15-49)	**Prevalence of HIV**
	רציף	תושבים שחיים באזור עירוני (% מכלל האוכלוסייה)	**Urban Population**
	רציף	מספר מקרי ההתאבדויות בשנה לכל 100,000 אנשים באוכלוסייה	**Suicide Mortality Rate**
The United Nations Statistics Division	קטגוריאלי	יבשת	**Continent**
Climatic Research Unit	רציף	טמפרטורה שנתית ממוצעת (צלזיוס)	**Temperature**
World Population Review	רציף	שיעור הפריון (מספר הילדים הממוצע שנשים בגיל הפוריות יולדות במדינה)	**Fertility Rate**

נספח 2 – ניתוח ראשוני של המשתנים הרציפים

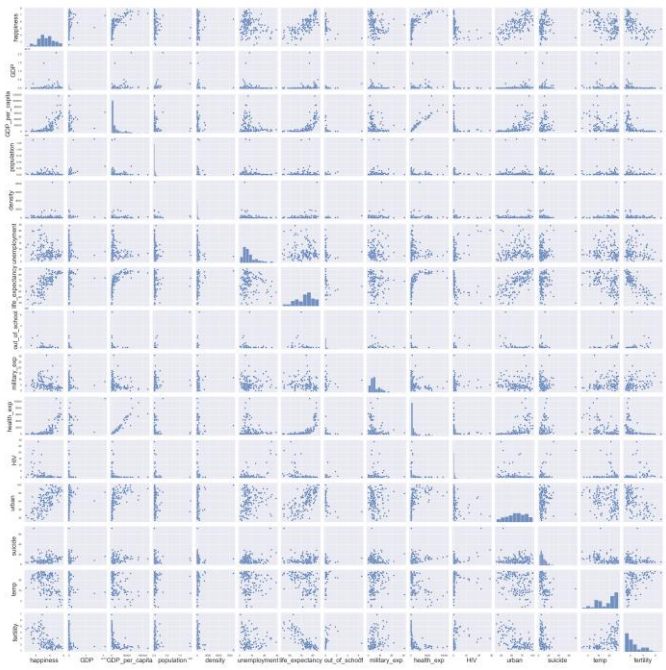
	Happiness Score	GDP	GDP per Capita	Total Population	Density of Population	Unemployment	Life Expectancy
count	151	148	148	151	133	150	151
mean	5.47	5.62E+11	1.45E+04	5.00E+07	236.53	8.24	72.85
std	1.12	2.17E+12	2.04E+04	1.65E+08	765.83	5.87	7.53
min	2.57	1.22E+09	2.34E+02	3.66E+05	2.18	0.33	53.68
25%	4.70	1.56E+10	1.84E+03	5.07E+06	41.93	4.35	67.38
50%	5.51	5.85E+10	5.14E+03	1.14E+07	83.93	6.41	74.35
75%	6.23	3.44E+11	1.81E+04	3.74E+07	199.64	10.90	78.09
max	7.81	2.09E+13	1.16E+05	1.41E+09	8322.69	29.22	85.39

	Children Out of School	Military Expenditure	Health Expenditure	Prevalence of HIV	Urban Population	Suicide Mortality Rate	Temperature	Fertility Rate
count	44	137	145	118	150	148	148	133
mean	3.54E+05	5.95	1264.20	1.79	61.01	9.48	17.26	2.60
std	1.05E+06	4.63	2021.32	4.40	22.19	7.75	8.56	1.27
min	26.00	0.00	19.85	0.10	13.71	1.60	-5.35	1.10
25%	1.46E+04	2.93	71.47	0.20	43.17	5.05	9.79	1.60
50%	4.78E+04	4.64	388.39	0.30	62.15	7.55	20.08	2.10
75%	2.29E+05	8.25	1342.07	1.10	80.45	11.83	24.83	3.40
max	6.55E+06	30.80	10921.01	28.60	100.00	72.40	28.29	6.90

נספח 3 – תרשימי צפיפות והיסטוגרמה



נספח 4 – תרשים Pairplot



נספח 5 – רגרסיה לינארית לתחזית HIV

		Beta	Std. Error	Sig
1	(Constant)	-2.818	.573	.000
	unemployment	.313	.055	.000
	suicide	.220	.041	.000